INTRODUCTION TO STATA

PART I

INTRODUCTION

Background

This handbook is designed to introduce **STATA for Windows XP**. It assumes familiarity with Microsoft Windows and standard windows-based office productivity software such as word processing and spreadsheets.

STATA is a popular and comprehensive data analysis package containing a multitude of features designed to facilitate the execution of a wide range of statistical analyses. It was developed in 1985 and is used world wide to aid research in economics, sociology, political science and epidemiology – STATA is short for Statistical Data Analysis and is well suited to; Data Management, Statistical analysis, Graphics, Simulations and Custom Programming.

STATA is predominantly a command driven package, however the majority of functions can be performed using drop down menus. The commands are more complicated to use than the menus, however they are more flexible having options that the menus do not, and once mastered often prove to be much more efficient. It should be noted that if a drop down menu is used the corresponding command will also be given. These notes will explain both procedures; it is up to the user to choose which they use.

This practical uses a set of data from a cross-sectional survey of respiratory function and dust levels amongst foundry workers. The object of the survey data is to determine whether the dust levels found in the foundries have any effect on the respiratory function.

When required, the data (in the form of an excel file and a .dta STATA data file) for this session can be found in the:

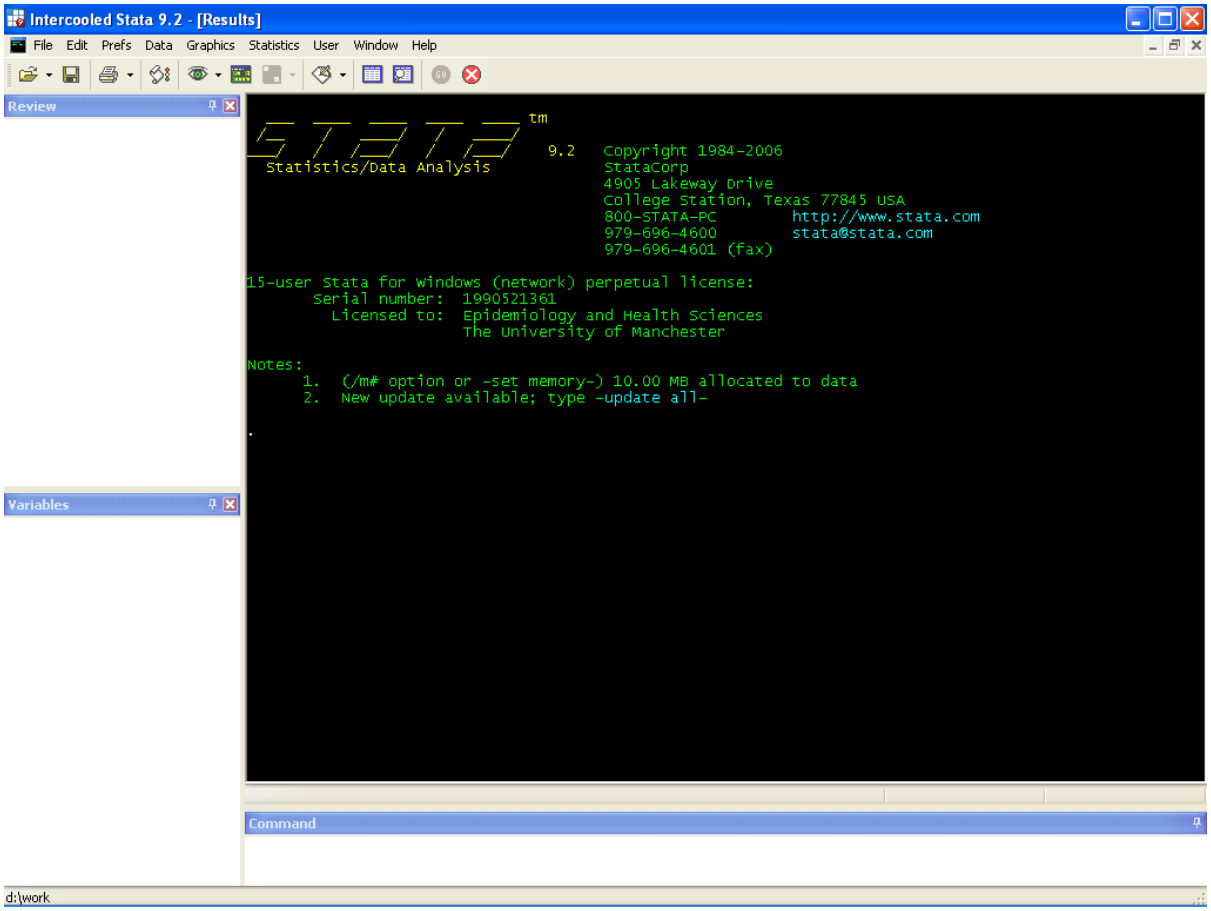
Shared Data area (found on the desktop) > mhs > Health Methodology Course Data

Starting STATA

After logging on to Windows XP, the user will be presented with a screen containing a number of different icons. Start STATA by clicking the **Start** button then selecting

All Programs > **Site Licensed Applications** > **Statistics** > **STATA V92**

Then a blank **STATA** screen will appear (shown below).



Window Orientation

The STATA screen above is the traditional layout and contains for windows.

Command Window – All STATA commands are typed and executed here

Results Window – Lists the output requested by the commands

Variables Window – Lists the variable names and variable labels in the current data set open in STATA. By clicking on a variable with the left mouse button in this window, the variable will appear in the command window

Review Window – Lists all previously used commands. As with the variable window, a command can be inserted into the command window by clicking on the review window command.

The standard window set up is as above, however this can be changed to suit the user and saved by clicking

Prefs > Manage Preferences > Save Preferences > New Preferences Set.

- The window sizes can be changed by clicking and holding the left mouse button on the edge of the window and then dragging to the required size.
- The results window font can be altered by right clicking on the results window, followed by font.

Command Structure

All STATA commands follow a common structure, below is a simplified version plus description which should help when formulating your own commands.

```
[by varlist:] command [varlist] [if] [in] [weight] [, options]
```

The command itself is the only compulsory element. Everything that is surrounded by a [] is considered to be an added option which is dependent on the analysis and methodology being used.

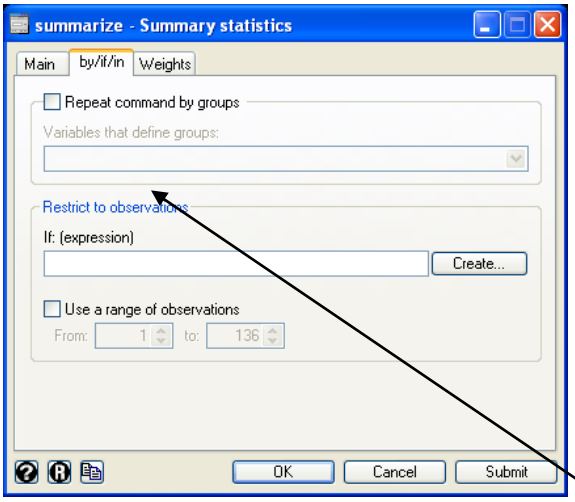
The Help Menu

All STATA commands come with a useful help file that explains the command fully along with the many options that can be applied to the command. The appropriate help file can be located in two ways. In the case where the command is already known to you then click **Help > Contents** and insert the command in the box provided. The appropriate help file will then appear in a separate window. The same result will occur by typing the command `help` followed by the command you are looking for.

Alternatively, if you do not know the command name of the analysis that you are looking for then **Help > Search** followed by an appropriate Key word will produce all the STATA files that contain this key word. A list of possibilities will appear in the results window and by clicking on the blue writing the corresponding help file will appear in a window. As with `help`, typing `search` followed by the key word will produce the same result.

Selecting a Subset of the Data

In addition to analysing the full set of data, you may want to analyse a subset. If, for example**,** you want to perform an analysis on Males only, in any menu driven command there should be a tab labelled **by/if/in.** Click on this tab and a window similar to the one shown here should appear.



Here you can choose one of three options to reduce your analysis to a subset.

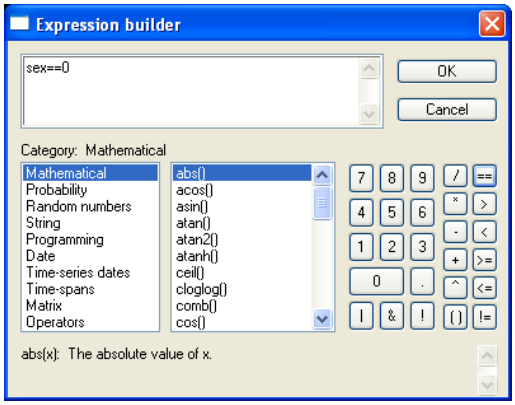
The by option repeats the analysis on groups of data. For example, the analysis can be repeated for males and females separately, to do so the variable representing gender (sex) should be placed here. In terms of a command the `by` command is placed before the analysis command you wish to perform (note STATA often requires that you sort the data by the grouping variables), for example to give the summary statistics of age for both males and females separately first sort the data then perform the analysis,

```
sort sex
by sex: summarize age
```

or alternatively use the `bysort` command;

```
bysort sex: summarize age
```

The `if` expression is used to restrict the analysis to a specific subset of the data, by clicking **create** an expression window will appear to allow you to restrict the analysis. For example, if we wish to perform the analysis on the males only, the expression **`sex==0`** is inserted here.



Click OK and follow the analysis through to its conclusion. In command format the `if` expression is added on to the end of the command

```
summarize age if sex==0
```

Note, that the `if` expression is not restricted to specific values (i.e. ==0), it can work with a variety of expression such as greater than or equal (>=), less than or equal (<=), etc. It is also possible to incorporate logic statements such as "and" & "or" using & , | respectively. For example, if we wished to summarize the variable for height for all those males who are 30 or older then the following command can be used.

```
summarize ht if sex==0 & age>=30
```

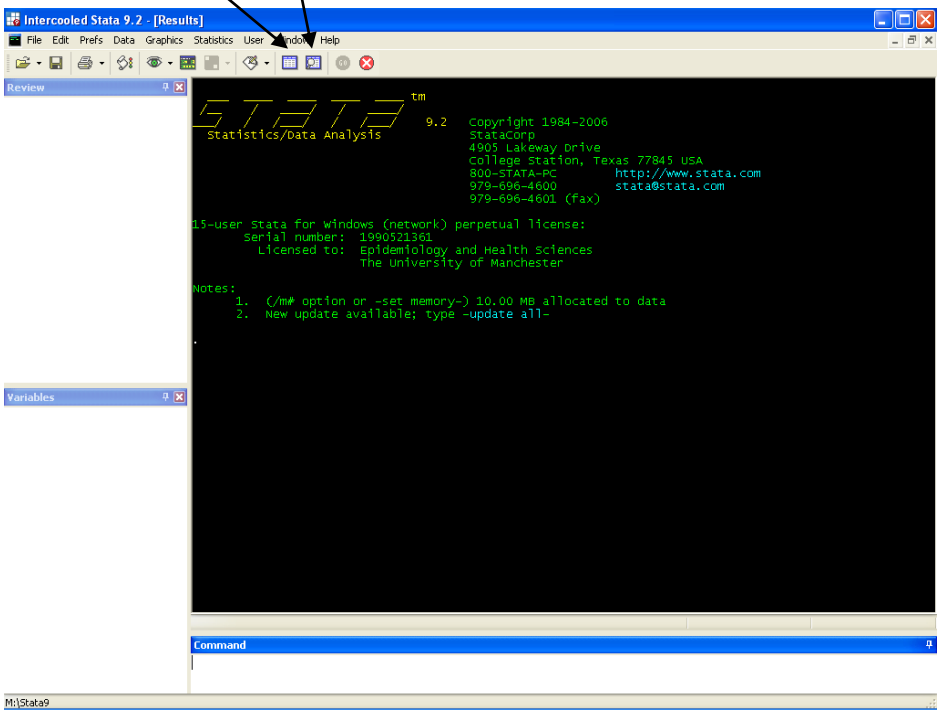
The final method to restrict the analysis is to a specific group of cases, For example the 1st fifty cases only. On the by/if/in window, click **Use a range of observations,** and then set to the range that you require. The command version uses `in` instead of `if` and define the range of observations with a `/` symbol. For example, to summarize the age variable for the first fifty cases the command would be.

.

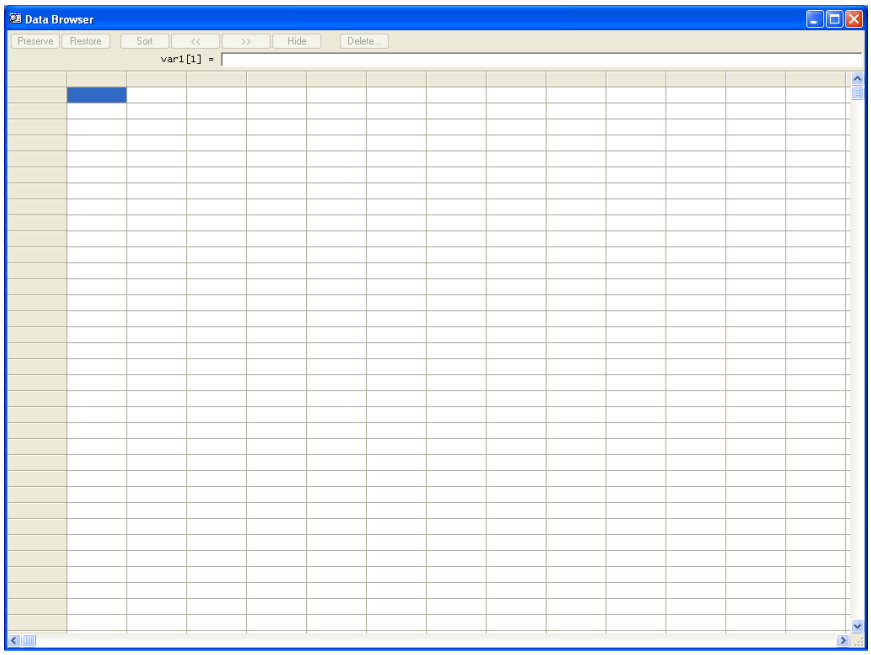
```
summarize age in 1/50
```

Inputting Data

In STATA the data screen can be accessed in two ways, through a data editor or a data browser. The difference being that unlike in data editor the data can not be altered in the data browser mode. To access either data editor or browser either click on the appropriate button on the menu at the top of the screen.



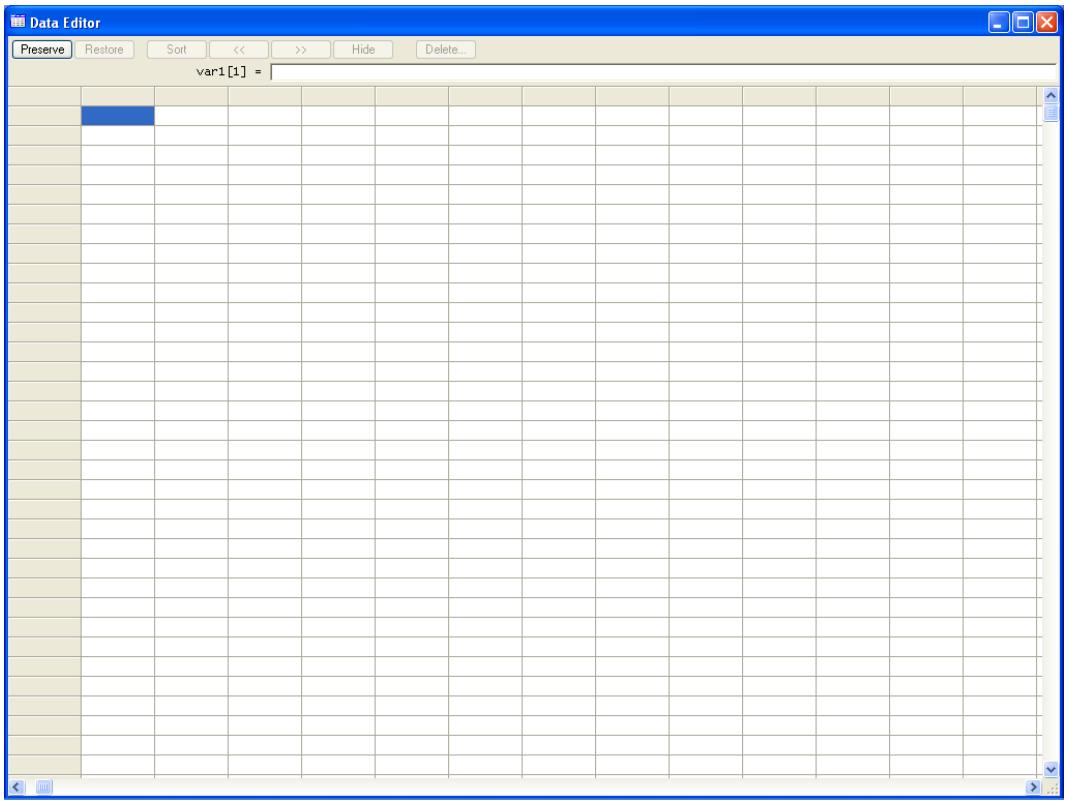
Or use the STATA commands `edit` or `browse`. The full data screen is below,



This is essentially in the same format as an Excel spreadsheet, with the columns representing the variables and the rows representing observations. As with Excel, data can be inserted here manually. However unlike Excel, each variable (column) can be defined so that they represent the correct structure of the data, e.g. continuous, categorical or string. A variable could be the answer to a question or any other piece of information recorded on each case. In STATA the data needs to be entered before you can define the variable, this is because STATA does not need the variable to be defined in order to perform the analysis (defining a variable especially in large datasets with many variables helps management and presentation of the data)

Entering Data

In the **Data Editor View** you will get the following blank screen



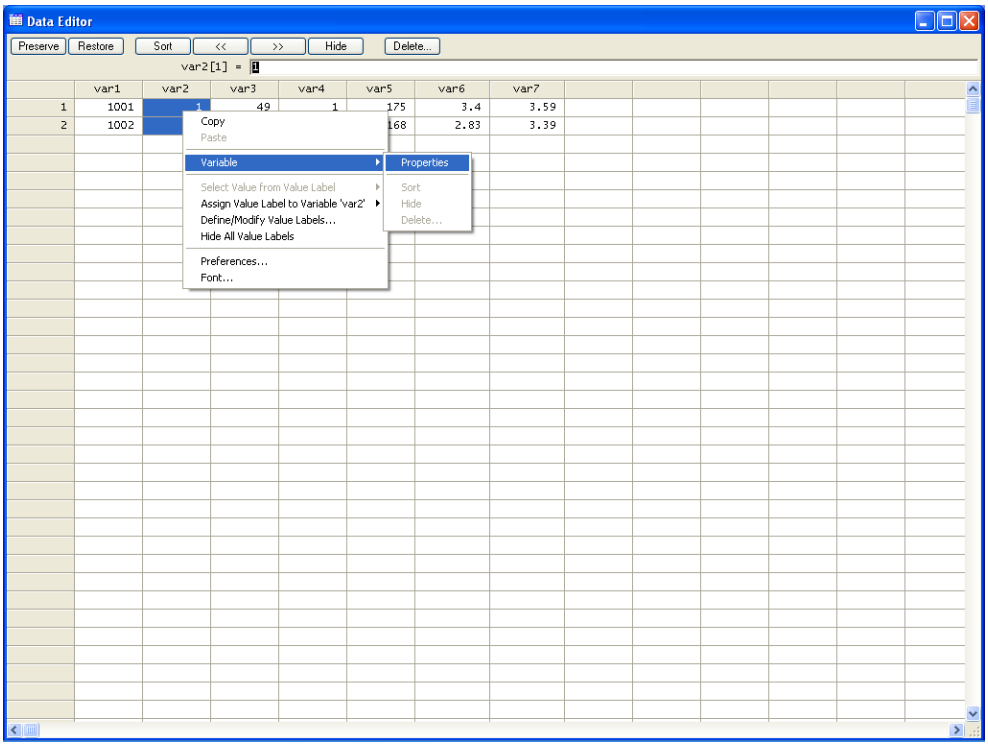
You can enter the data straight away as you would in a spreadsheet. To make an entry in a particular cell on the spreadsheet use the mouse to move the cursor to select that cell and type in the value. The value will appear in the cell. Click on the mouse or press enter to enter that value. Note, at this stage STATA assumes that all variables are numerical and any data entered not numerical will be rejected. Therefore, in the case of categorical data a word may represent a group, e.g. males or females. Assign a value to each category (0=males, 1=females) and insert the number, value labels can be assigned later. If incorrect data is entered, it can be overtyped or deleted.

<u>Exercise</u> The data below is from the foundry study for which you will enter the variable codes later. Enter the first couple of lines into the work sheet. If you leave a gap in any cell in the worksheet, **STATA** will put a dot (**.**) and treat it as missing data. At this stage do not insert the variable names at the top instead enter the data from the second row down (idno=1001). In each case **enter the numerical value** corresponding to appropriate characteristic as indicated in the first row at the top of each column, the corresponding value labels will be added shortly. For example, **enter 0 for Females and 1 for Males**

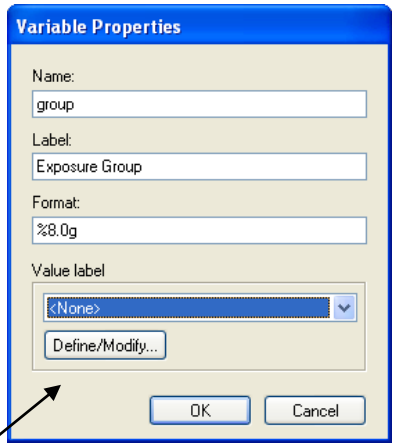
idno	group N=0, E=1	age	sex F=0, 1=1	ht	fevmeas	fevpred	fvcmeas	fvcpred	asthma N=0, Y=1	bron N=0, Y=1	smknow N=0, Y=1	smkever N=0, E=1, C=2	cigno	cigyrs	empyrs	respdust
1001	Exp	49	Female	175	3.40	3.59	4.49	4.45	No	No	Yes	Curr	20	31	23	1.71
1002	Exp.	46	Female	168	2.83	3.39	3.91	4.12	Yes	No	Yes	Curr	20	11	16	0.69
1003	Non	34	Female	180	3.93	4.26	4.80	5.14	No	No	No	Never			12	0.00
1004	Non	34	Male	180	4.01	4.25	4.57	5.12	No	No	Yes	Curr	25	16	12	0.00

Defining Variables – Variable & Value Labels

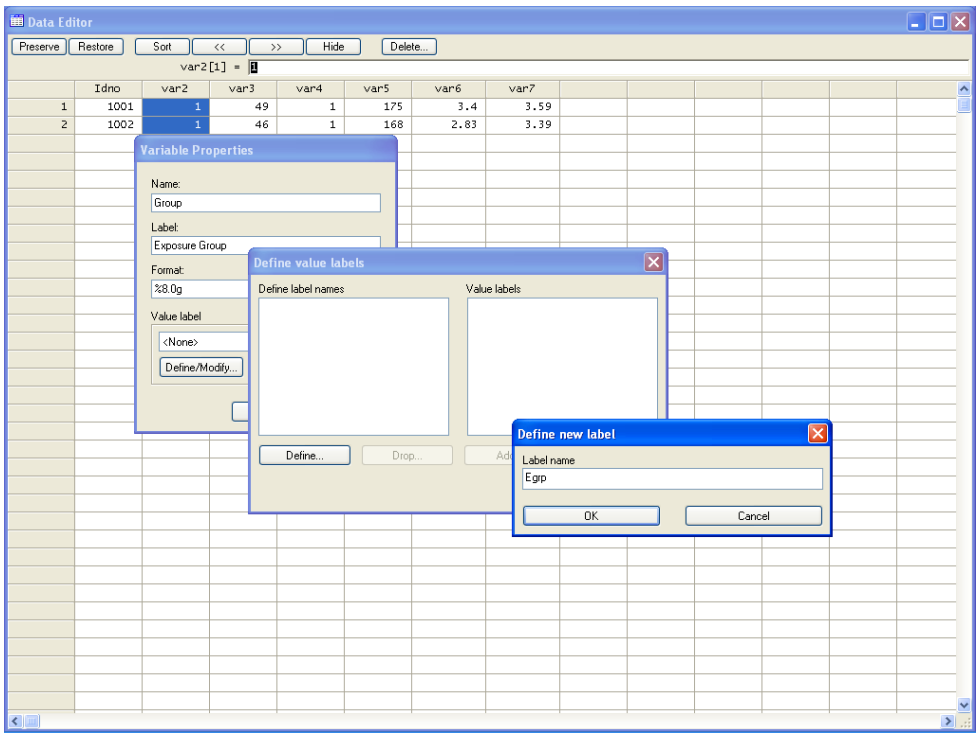
After entering the first few lines you are now ready to define the variables. Left click on the **Column header** > **Variable** > **Properties**



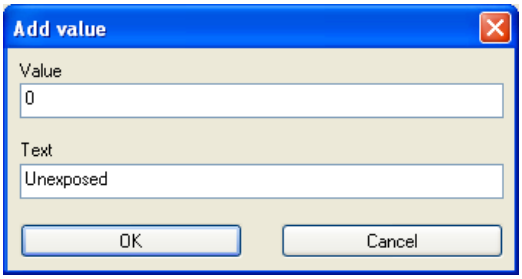
This results in the following window, for example open the second variable properties window.



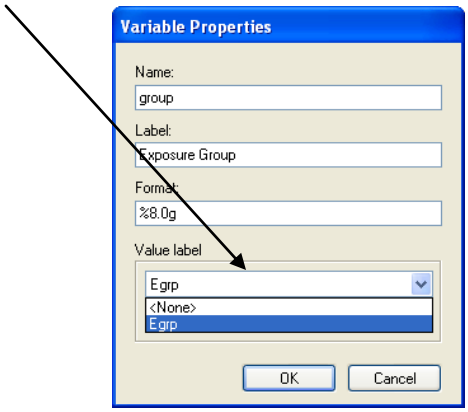
This window allows you to define the variables properties, such as its name, label and in the case of a categorical variable the value labels. In the case of a categorical variable, first a value label is defined and then it is assigned to a variable. This saves time if a dataset has more than one variable with the same label as there is no need to repeatedly define variable labels. Click the **Define/Modify** button followed by **Define** to bring up the following window.



First enter a **Label name** in the Define new label box and click ok. This can be anything, however, it is a good idea to keep it relevant as it may be required again. A new screen will then automatically appear.



Here insert an appropriate value and its corresponding text, for example the above variable, Exposure groups falls into two categories Unexposed (0) and Exposed (1). Insert one after the other clicking ok in between. Each time the settings will be automatically placed in the **Defining value labels** window. When the last category has been inserted click Cancel in the **Add value** box followed by Close, you will then return to the original window. At this point use the drop down arrow to select a value label.



The final step is to close the Data Editor window, at this point it will ask you to accept or reject the changes you have made click accept and you return to the main screen.

As with all procedures in STATA there is a corresponding set of commands that will do the exactly what has been discussed over the last few pages, these are automatically put into the **Results Window** when **Data Editor** is closed down.

The following commands in an alternative order can be seen, note the descriptions have been added later.

```
rename var2 group
```

Alters the variable name from var2 to Group

```
label var Group "Exposure Group"
```

Alters the variable label for Group to Exposure Group

```
label values group Egrp
```

Defines the label for the values of the variable Group as Egrp

```
label define Egrp 0 "Unexposed" 1 “Exposed”
```

Within the value label Egrp this defines the values in Group to represent Unexposed and Exposed

Exercise The table below lists the variables from the foundry study. Set-up the following variables

Variable Name	Description (Variable Label)	Type	Extras	Value Labels for each code
idno	Identification No	Numeric		
age		Numeric		
group	Exposure Group	Numeric	Labels	1 = Exposed to dust 0 = Unexposed
sex		Numeric	Labels	0 = female 1 = male
ht	Height in cms	Numeric		
fevmeas	Measured FEV	Numeric		
fevpred	Predicted FEV	Numeric		
fvcmeas	Measured FVC	Numeric		
fvcpred	Predicted FVC	Numeric		
asthma	Ever had asthma	Numeric	Labels	0 = No 1 = Yes 2 = Don't Know
bron	Ever had Bronchitis	Numeric	Labels	0 = No 1 = Yes 2 = Don't Know
smknow	Do you smoke now	Numeric	Labels	1 = Yes 0 = No
smkever	Have you ever smoked	Numeric	Labels	0 = No 1 = Ex smoker 2 = Current smoker
cigno	No of cigarettes per day	Numeric		
cigyrs	No of years smoked	Numeric		
empyrs	No of Years with company	Numeric		
respdust	Current exposure	Numeric		

Reviewing Variables

Once you have created all these variables, you can check they have been set up correctly. To do this click from the menu bar **Data > Describe Data > Describe Variables in Memory,** then either insert the variable you require or leave blank for all variables. The following screen should appear in the results window.

```
Contains data from P:\mbbxdmg2\My Notes\Statanotesdata.dta
  obs:            10
 vars:            17                          25 Sep 2006 14:44
 size:           380 (99.9% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
idno            int    %8.0g                  Identification No
group           byte   %8.0g       Egrp       Exposure Group
age             byte   %8.0g
sex             byte   %8.0g       sex
ht              int    %8.0g                  Height in cms
fevmeas         float  %9.0g                  Measured FEV
fevpred         float  %9.0g                  Predicted FEV
fvcmeas         float  %9.0g                  Measured FVC
fvcpred         float  %9.0g                  Predicted FVC
asthma          byte   %8.0g       Astbron    Ever had Asthma
bron            byte   %8.0g       Astbron    Ever had Bronchitis
smknow          byte   %8.0g       Astbron    Do you smoke now
smkever         byte   %8.0g       smkever    Have you ever smoked
cigno           byte   %8.0g                  No of cigarettes per day
cigyrs          byte   %8.0g                  No of years smoked
empyrs          byte   %8.0g                  No of years with the company
respdust        float  %9.0g                  Current Exposure
-------------------------------------------------------------------------------
Sorted by:
```

Note, that the same will appear when using the following command

```
describe
```

Also note, as with the drop down menu, by specifying no variables using the describe command all shall be included in the output.

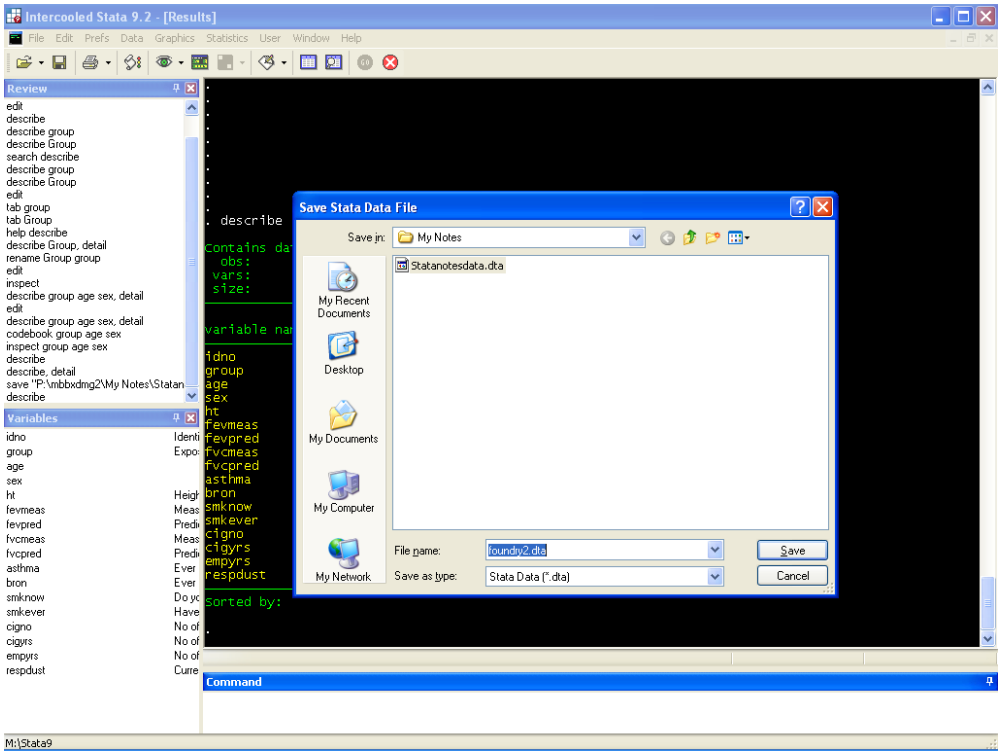
FILE MANAGEMENT

Saving a STATA File

Once you have entered some data you will want to save it to disk. It is good practice to save data at regular intervals during data entry just in case!

To save the data you have just entered, click on the **File** option at the top left corner of the screen and then on the **Save As...** sub-option.

The following screen will appear:



To save a copy of the current **STATA** file on your floppy disk, under **Drives:** click on ▼ in the **save in** window to generate a list of the drives.

Click on the up-arrow to move to the **3½ Floppy (A): / Memory stick etc**, drive and move the cursor to the **File name** window and enter a suitable name. By default STATA will add the file extension **.dat**. Finally, click on the **Save** button. It will help to identify the file as a STATA datafile if the file extension **.dat** is used.

Alternatively use the STATA command

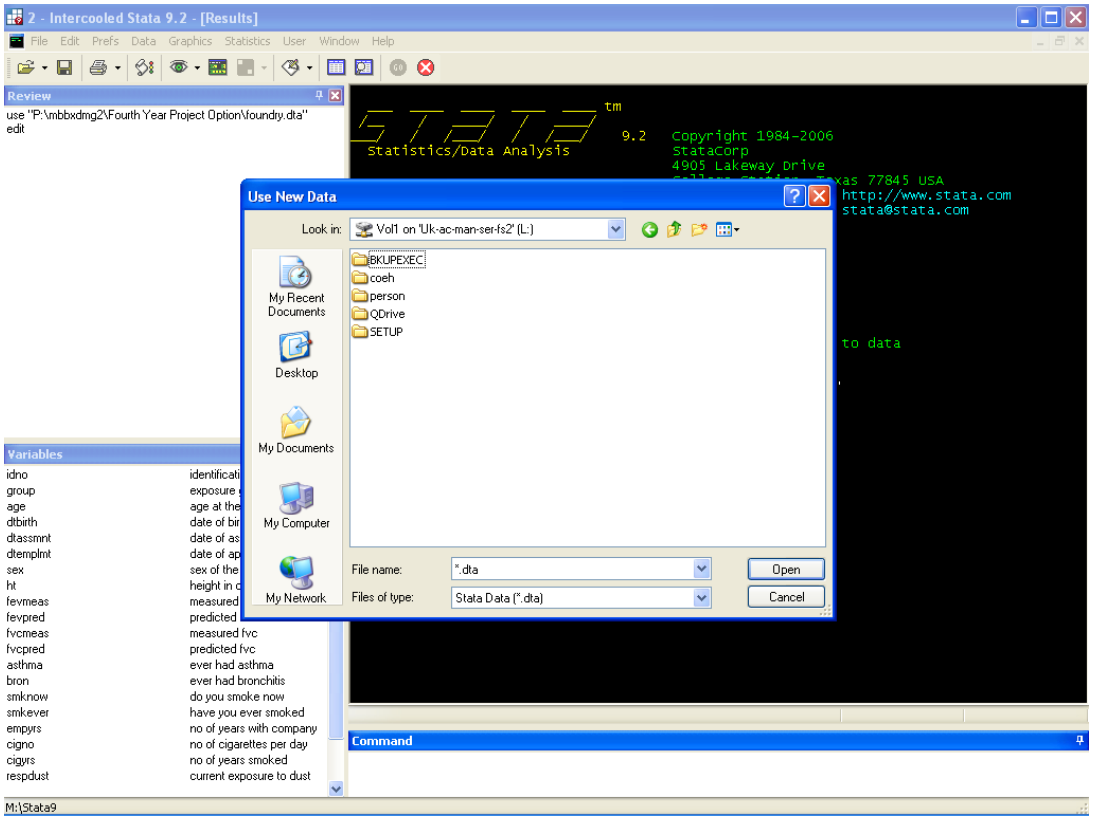
```
     save "Filename including Full directory Route"
e.g. save "C:\Notes\Stata Tutorial\mydata.dta"
```

Backing Up Your Data

It is good practice to save data on different disks and also several names as data entry progresses (e.g. **mydata1 mydata2** etc). To make a backup copy of your data, repeat the **Save Data As** procedure.

Retrieving Data Files

Retrieving STATA File is essentially the reverse of this process. Click on the **File** option at the top of the screen, then the **Open** option. The following screen will appear. Then select the required file from the window.



We can also open a data file using the command.

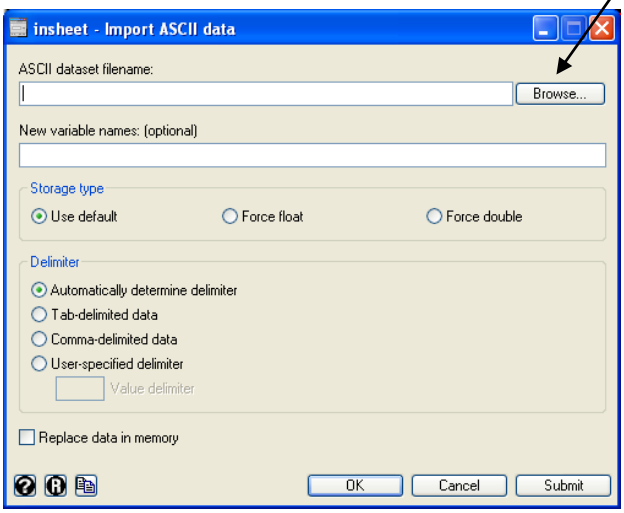
```
use "Full File Directory"
```

Note, you cannot have two files open in the same STATA window, therefore if another file is currently open the option `, clear` needs to be added onto the end of the use command.

Reading An Excel File Into STATA

Often data may be already stored in another data format. STATA has the ability to read many of these. For example you can retrieve an Excel file into STATA. If you put the variable names in the first row of your spreadsheet, they can be copied as variable names in the STATA file. Unlike StatsDirect, STATA is only able to read a single work sheet it cannot read a complete work book with several sheets. In order that STATA can read it, the Excel file needs to be saved in **CSV format**.

The data from the foundry study is saved in a spreadsheet **Shared Data area (found on the desktop) > mhs > Health Methodology Course Data**. The names of the variables have been entered in the first row. You may wish to check this by going to EXCEL. The procedure for retrieving the data from EXCEL is similar to retrieving an STATA data file. Click on the **File** option at the top of the screen, then **import** followed by **ASCII data created by a spreadsheet** option so that the screen bellow appears. At this point click the **browse** button and locate the file required.



Then press **OK**. You will get output in the results window reporting the number of variables and observations, the variable window will now contain a list of variables. Note that at this point the data is in a raw format, which means the variable and value properties will need to be set up, please see previous section **Defining Variables**.

The command;

```
insheet using "filename"
```

will perform the same action.

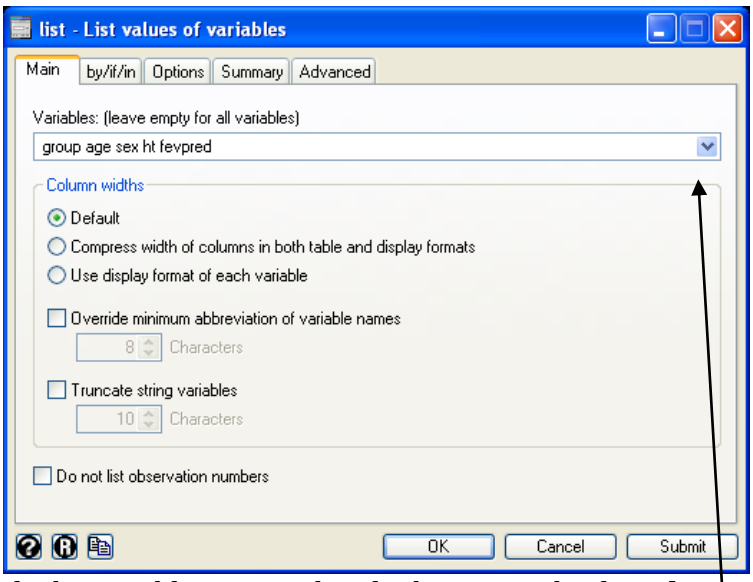
Having read data from an excel spreadsheet it is important to check what has been read in. For example if a column on the spreadsheet contained a mix of numeric and string data (besides the variable name at the top) either one or the other may be set to missing.

INITIAL DATA CHECKING

For the next stage you need to retrieve the data file **foundry** which contains the data with variable and value labels. As before click **Shared Data** icon on the desktop then **mhs, Health Methodology Course Data** then click the link **foundry.dta** for the dataset followed by **open.**

Case Summaries

With any data set it is extremely **IMPORTANT** that you check the data entered as carefully as possible. One way you can do this is, to list case. To do this, you use **Data**, **Describe** and then **List.**



In this window, include the variables you wish to look at using the drop **down menu**, if you wish to look at all the variables leave this blank, then click **ok**. The facility allows you to look at a column or columns separately from the rest of the data. The following output appears, note only the first 10 are shown here.

It is then easy to see any potential errors e.g. if there was "never" in ever smoked and "yes" in do you smoke now, there has been an error made. The left-hand side column is the case number.

The associated command for this is `list`, either followed by a variable list or left blank.

```
     +-----------------------------------------+
     |     group   age      sex    ht   fevpred |
     |-----------------------------------------|
  1. |  exposure    49   female   175      3.59 |
  2. |  exposure    46   female   168      3.39 |
  3. |  exposure    34   female   180      4.26 |
  4. | unexposed    34     male   180      4.25 |
  5. | unexposed    29     male   183      4.52 |
     |-----------------------------------------|
  6. |  exposure    43     male   174      3.73 |
  7. |  exposure    27     male   180      4.45 |
  8. |  exposure    59   female   167      2.97 |
  9. |  exposure    29   female   175      4.18 |
 10. |  exposure    31   female   177      4.21 |
     |-----------------------------------------|
```

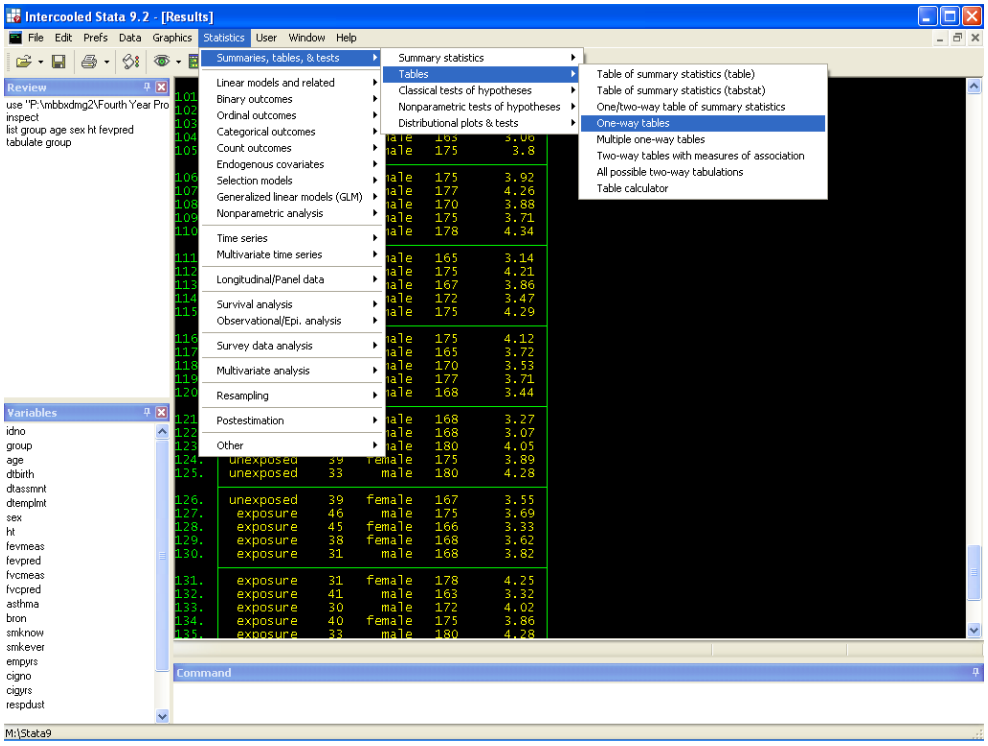
DESCRIPTIVE STATISTICS

The first step in data analysis is to generate descriptive statistics. This will give us a feel for the data. It will also help us identify any inconsistencies that there may be in the data. This is sometimes called data cleaning. Techniques that are commonly used to do this include:

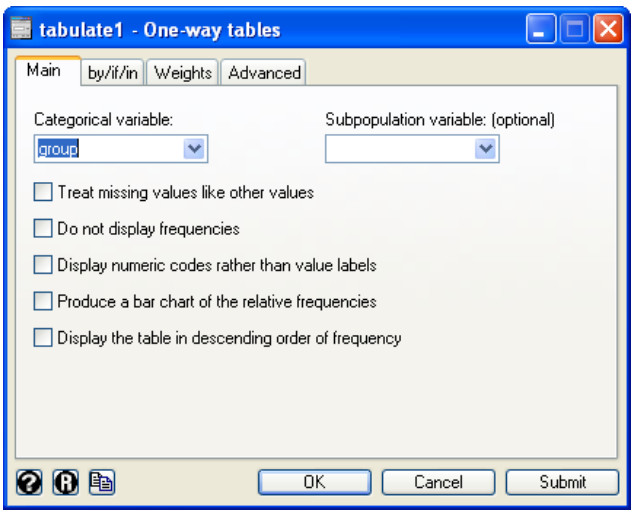
- Frequency Analyses
- Descriptive Statistics
- Cross-tabulations
- Plots

Frequency Tables

A basic way to check for data errors is by carrying out a frequencies analysis on variables, to do this click on the following sequence **Statistics > summaries, tables & tests > tables > One-way tables** as shown below.



A menu window will then appear, in which enter a categorical variable. In this case we have chosen exposure group to demonstrate.



Click on **OK**. A frequency table will be given in the results screen, the example gives:

```
  exposure group |      Freq.     Percent        Cum.
-----------------+-----------------------------------
       unexposed |         63       46.32       46.32
exposure to dust |         73       53.68      100.00
-----------------+-----------------------------------
           Total |        136      100.00
```

Frequency tables can be copied into word processing documents by highlighting the table and selecting **Edit** then **Copy.** To place in the word processing document, use **Edit** and **Paste**.

The following STATA command produces the same result

```
tabulate group
```

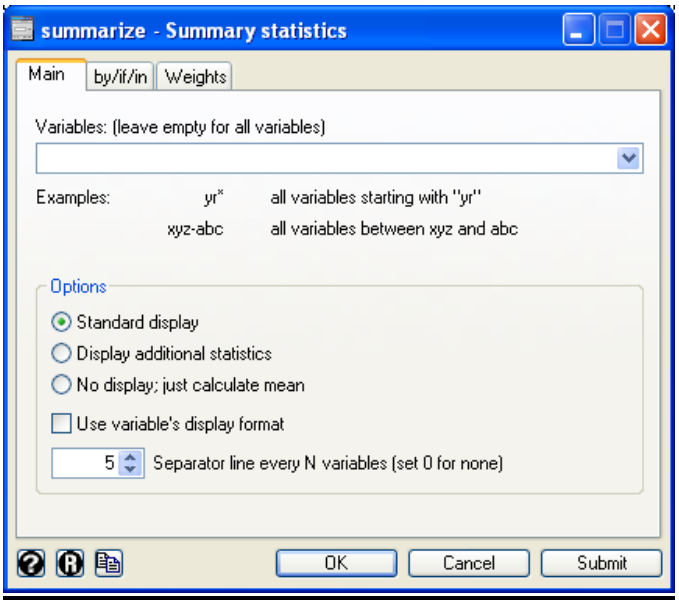
<u>Exercise</u> Using the frequencies options find out

- what proportion of the foundry workers were exposed to dust?
- what proportions had ever suffered from bronchitis?
- what proportion had ever smoked?
- what proportion smoked more than 40 cigarettes per day?

Descriptives

Descriptive statistics can be calculated for quantitative data in STATA by using the summarize commands. To use this click **Statistics** > **Summaries, tables & tests** > **Summary statistics** > **Summary statistics** the window bellow will then appear. Insert the variables of interest into the **Variables** box by the drop down tab on the right-hand side. As with the frequencies command we can obtain descriptive statistics for several variables at once. The standard display will give the mean, standard deviation, minimum, maximum and the number of observations. If the button

Display additional statistics is pressed then a further descriptive statistics such as median, variance, percentile point, etc will be calculated..



To obtain through the commands use;

```
summarize [varname]
```

to add on the extra details add `, detail` to the command.

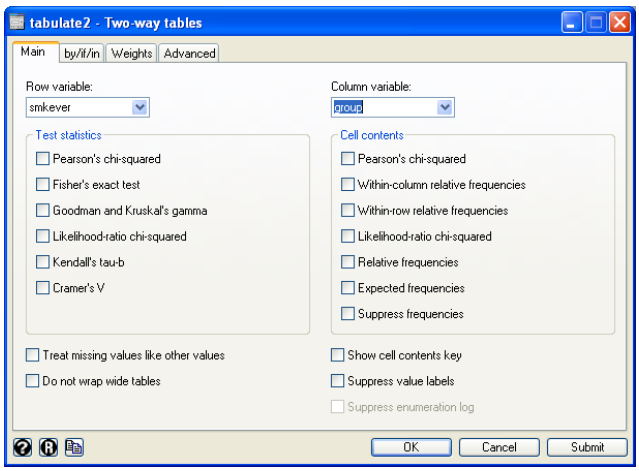
<u>Exercise</u> Use the **descriptive** procedure to determine

- the current mean exposure to dust per day
- the mean number of cigarettes smoked per day

Cross-tabulation

To examine the relationship between two categorical variables, a two way Frequency Table can be used. This is called a cross-tabulation. Click **Statistics** > **Summaries, tables & tests** > **Tables** > **Two-way tables with measures of associations** the screen below appears. Suppose we wished to examine how smoking status related to exposure. We could examine this by a cross-tabulation of the variables **group** and **smkever**.

Select the smoking status variable **smkever** in the row variable box by the drop down tab to make this the row variable. Then by the same method select **group** labelled **Exposure Group** in the column variable list by to finally press **OK**



You may notice several methods of testing the association between variables, these will be covered further in a later section. The following result appears when the two frequency table has been completed.

```
    have you |     exposure group
 ever smoked | unexposed  exposure  |     Total
-------------+----------------------+----------
       never |        24         20 |        44
   ex smoker |        19         19 |        38
curr. smoker |        20         34 |        54
-------------+----------------------+----------
       Total |        63         73 |       136
```

The same command used when calculating a one way frequency table is used here to calculate a two way table, the only difference is the inclusion of a second variable. So to get the output above type;

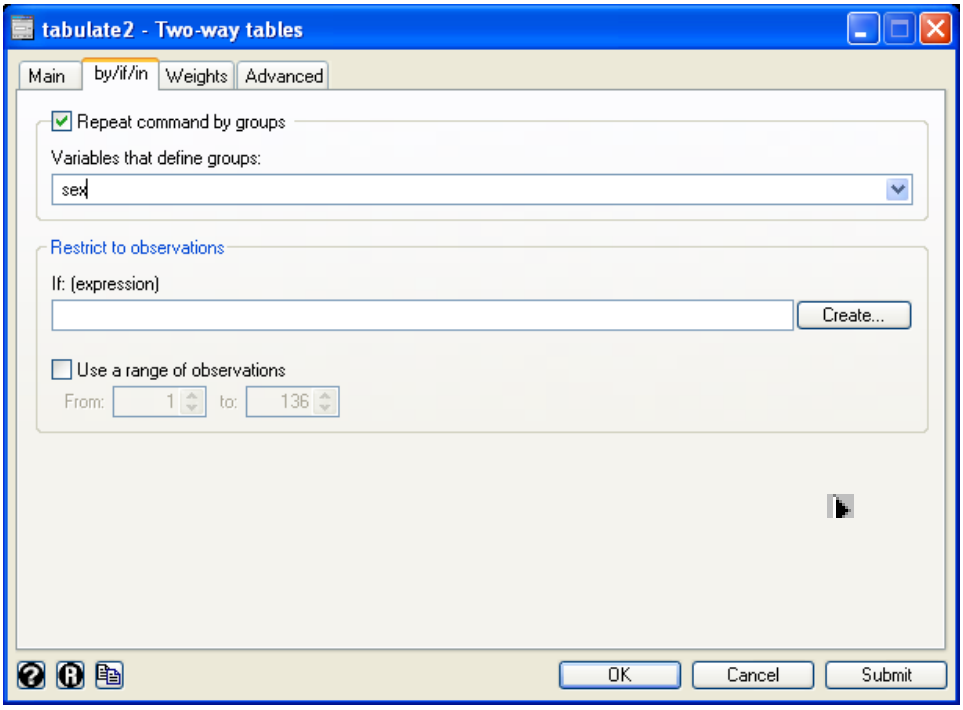
```
tabulate smkever group
```

Two way frequency tables are more informative if they include percentages. Adding percentages to the table cannot be done through the drop down menu however it can be achieved through the commands. By adding `row` and/or `column` on the end of the command, row percentages and/or column percentages respectively will be included in the output. Including both sets of percentages can make the output confusing, therefore it may be beneficial to do the separately. For the table above column percentages are the most useful as they will allow us to compare the smoking status of non-exposed and exposed subjects. By writing the following command we obtain the following output.

```
 tabulate smkever group, column

    have you |     exposure group
 ever smoked | unexposed  exposure  |      Total
-------------+----------------------+----------
       never |        24         20 |         44
             |     38.10      27.40 |      32.35
-------------+----------------------+----------
   ex smoker |        19         19 |         38
             |     30.16      26.03 |      27.94
-------------+----------------------+----------
curr. smoker |        20         34 |         54
             |     31.75      46.58 |      39.71
-------------+----------------------+----------
       Total |        63         73 |        136
             |    100.00     100.00 |     100.00
```

Three-way tables

You may need to do comparisons on three variables. In STATA there is no specific command or menu that does this for you, however it is possible to get the same result by using the by option. To do this follow the same procedure as for the two-way table show previously, however when the window appears click the **by/if/in** tab.



Tick the box marked **Repeat command by groups** and include the variable of choice in the **Variables that define groups** box. The same can be achieved by adding `bysort sex:` onto the from of the `tabulate` command. The output following this appropriate command, is obtained;

```
bysort sex: tabulate smkever group
```

```
> sex = male
    have you |    exposure group
 ever smoked | unexposed  exposure  |     Total
-------------+----------------------+----------
       never |        14          6 |        20
   ex smoker |         7          7 |        14
curr. smoker |        12         17 |        29
-------------+----------------------+----------
       Total |        33         30 |        63

> sex = female
    have you |    exposure group
 ever smoked | unexposed  exposure  |     Total
-------------+----------------------+----------
       never |        10         14 |        24
   ex smoker |        12         12 |        24
curr. smoker |         8         17 |        25
-------------+----------------------+----------
       Total |        30         43 |        73
```

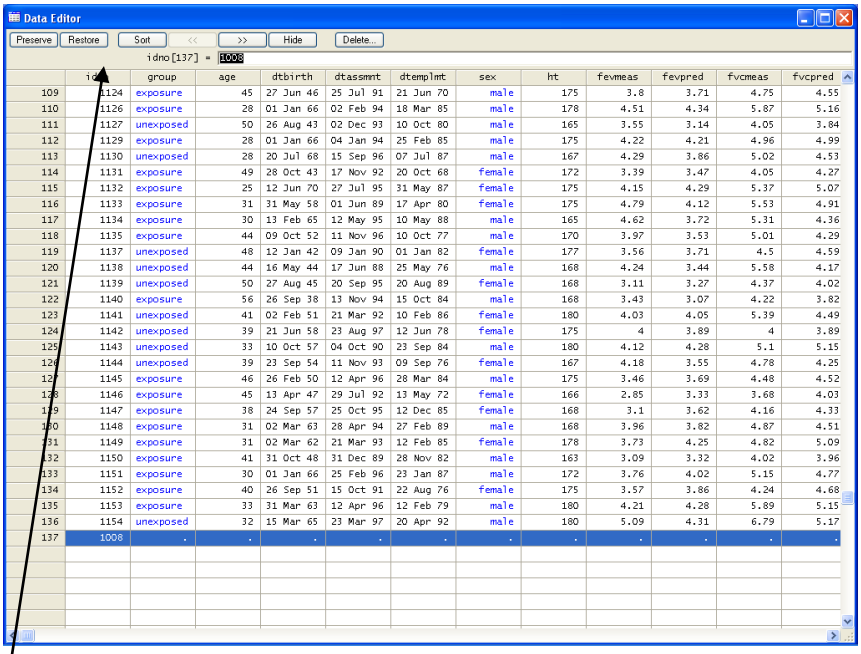
EDITING AND MODIFYING THE DATASET

Having done some preliminary analysis we may need to change the data. There are some useful functions for modifying data files. Firstly, note that once a change to the data set has been performed it will be lost and cannot be undone. To combat this if you are unsure of what you are going to do then you can type `preserve` before proceeding with any commands and then if you decided that you wish to undo what you have done typing `restore` will return the dataset to the condition it was in when you typed `preserve`.

Inserting Data

You may have noticed that idno 1008 was missing.

To insert it, either enter the **Data Editor** and insert case (along with all its details) into the first new blank row after idno 1154 (in this case row 137), shown below.



	idno	group	age	dtbirth	dtassmnt	dtemplmt	sex	ht	fevmeas	fevpred	fvcmeas	fvcpred
109	1124	exposure	45	27 Jun 46	25 Jul 91	21 Jun 70	male	175	3.8	3.71	4.75	4.55
110	1126	exposure	28	01 Jan 66	02 Feb 94	18 Mar 85	male	178	4.51	4.34	5.87	5.16
111	1127	unexposed	50	26 Aug 43	02 Dec 93	10 Oct 80	male	165	3.55	3.14	4.05	3.84
112	1129	exposure	28	01 Jan 66	04 Jan 94	25 Feb 85	male	175	4.22	4.21	4.96	4.99
113	1130	unexposed	28	20 Jul 68	15 Sep 96	07 Jul 87	male	167	4.29	3.86	5.02	4.53
114	1131	exposure	49	28 Oct 43	17 Nov 92	20 Oct 68	female	172	3.39	3.47	4.05	4.27
115	1132	exposure	25	12 Jun 70	27 Jul 95	31 May 87	female	175	4.15	4.29	5.37	5.07
116	1133	exposure	31	31 May 58	01 Jun 89	17 Apr 80	female	175	4.79	4.12	5.53	4.91
117	1134	exposure	30	13 Feb 65	12 May 95	10 May 88	male	165	4.62	3.72	5.31	4.36
118	1135	exposure	44	09 Oct 52	11 Nov 96	10 Oct 77	male	170	3.97	3.53	5.01	4.29
119	1137	unexposed	48	12 Jan 42	09 Jan 90	01 Jan 82	female	177	3.56	3.71	4.5	4.59
120	1138	unexposed	44	16 May 44	17 Jun 88	25 May 76	male	168	4.24	3.44	5.58	4.17
121	1139	unexposed	50	27 Aug 45	20 Sep 95	20 Aug 89	female	168	3.11	3.27	4.37	4.02
122	1140	exposure	56	26 Sep 38	13 Nov 94	15 Oct 84	male	168	3.43	3.07	4.22	3.82
123	1141	unexposed	41	02 Feb 51	21 Mar 92	10 Feb 86	female	180	4.03	4.05	5.39	4.49
124	1142	unexposed	39	21 Jun 58	23 Aug 97	12 Jun 78	female	175	4	3.89	4	3.89
125	1143	unexposed	33	10 Oct 57	04 Oct 90	23 Sep 84	male	180	4.12	4.28	5.1	5.15
126	1144	unexposed	39	23 Sep 54	11 Nov 93	09 Sep 76	female	167	4.18	3.55	4.78	4.25
127	1145	exposure	46	26 Feb 50	12 Apr 96	28 Mar 84	male	175	3.46	3.69	4.48	4.52
128	1146	exposure	45	13 Apr 47	29 Jul 92	13 May 72	female	166	2.85	3.33	3.68	4.03
129	1147	exposure	38	24 Sep 57	25 Oct 95	12 Dec 85	female	168	3.1	3.62	4.16	4.33
130	1148	exposure	31	02 Mar 63	28 Apr 94	27 Feb 89	male	168	3.96	3.82	4.87	4.51
131	1149	exposure	31	02 Mar 62	21 Mar 93	12 Feb 85	female	178	3.73	4.25	4.82	5.09
132	1150	exposure	41	31 Oct 48	31 Dec 89	28 Nov 82	male	163	3.09	3.32	4.02	3.96
133	1151	exposure	30	01 Jan 66	25 Feb 96	23 Jan 87	male	172	3.76	4.02	5.15	4.77
134	1152	exposure	40	26 Sep 51	15 Oct 91	22 Aug 76	female	175	3.57	3.86	4.24	4.68
135	1153	exposure	33	31 Mar 63	12 Apr 96	12 Feb 79	male	180	4.21	4.28	5.89	5.15
136	1154	unexposed	32	15 Mar 65	23 Mar 97	20 Apr 92	male	180	5.09	4.31	6.79	5.17
137	1008	.	.	.	.	.	.	.	.	.	.	.

By clicking the **sort** button the data is then ordered by the first variable (idno).

The same is can be achieved by the commands `set` and `replace`.

```
set obs 137
replace idno = 1008 in 137
replace group = 1 in 137
replace age = 38 in 137
replace sex = 1 in 137
(etc)
sort idno
```

note, the `in 137` specifies the exact row the value should be placed in, by not including this the entire variable may be changed. Therefore it is often useful to use the command `preserve` before this procedure in order that any large mistake and the data set can be returned back to its original form using `restore`.

You can insert the following case (idno 1008) in the blank line

Variable	Value	Variable	Value
Idno	1008	Asthma	0
Group	1	Bron	0
Sex	1	Smknow	1
Ht	180	Smkever	2
Fevmeas	4.01	Cigno	30
Fevpred	4.45	Cigsyrs	20
Fvcmeas	4.90	Empyrs	10
Fvcpred	5.30	Respdust	2.04
Age	38		

Deleting A Case

To delete a case, click on its number on the left of the **Data Editor** screen to highlight the row containing the case. Press the **Delete** button at the top of the window followed by **Delete observation [case no]** and then click **ok**. The case will then be removed and the rest will move up to fill the gap.

Alternatively, use the command `drop` with the `if` constraint to specify which case to delete. For example if we wished to delete the first case (1001) the command would be;

```
drop if idno == 1001
```

Exercise Delete case no 1008

Deleting A Variable

To delete a variable, click on its name at the top of the **Data Editor** to highlight the column containing the variable. Then press the **Delete** button followed by **Delete variable [varname]** and **ok**. The variable is deleted and the variables to the right move to the left to fill the gap.

The command `drop` can also be used in this situation, this time do not use the `if` constraint and instead just define the variable name to be deleted. For example if we wished to delete age;

```
drop age
```

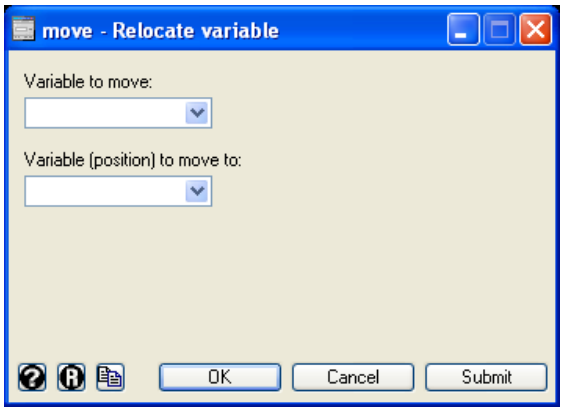
Deleting An Entry In An Individual Cell

To delete an entry in an individual cell, open the **Data Editor** screen click the cell you wish to delete and either press the **delete button** on the key board followed by **enter** or click the delete button at the top of the window followed by **Delete all 1 obs, where [varname]==[outcome].** Alternatively combine the two previous commands in order to specifically delete one cell. For example, if we wished to delete observation no 1007's age then the following command will be employed,

```
drop age if idno==1007
```

Moving A Variable

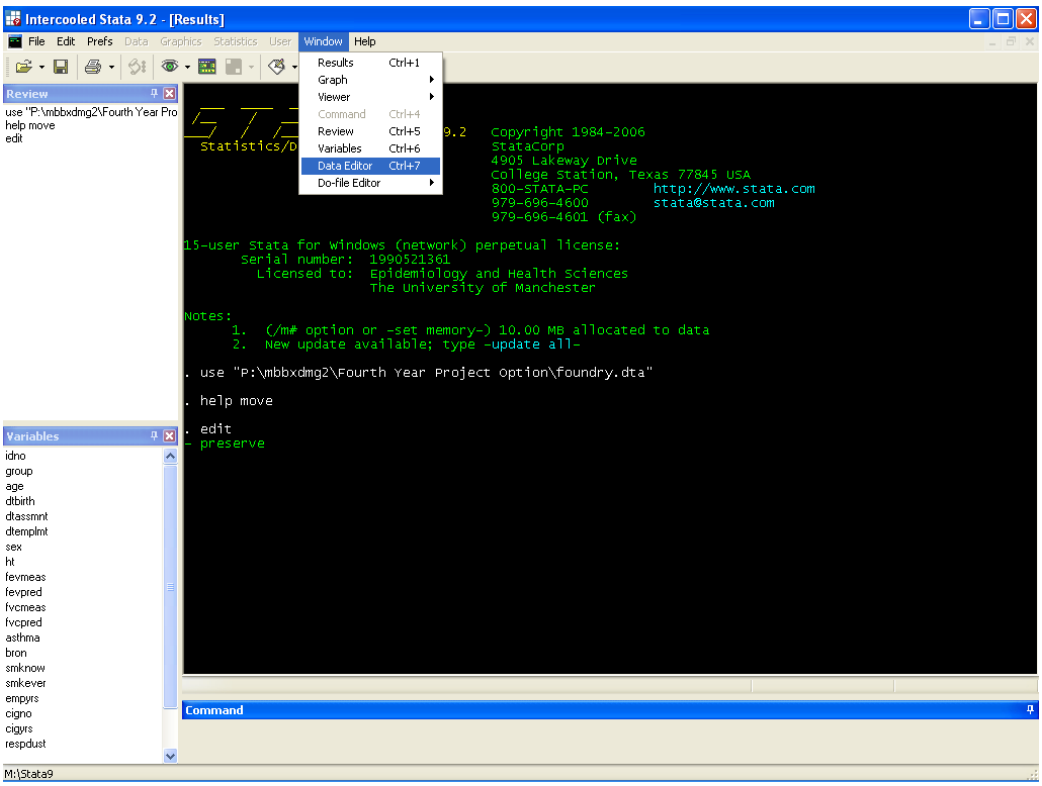
Click, **Data > Variable Utilities > Relocate variable** to get the following window open, then insert the variable to move and the Variable position to move to (note that it will be positioned behind this variable).



The command `move` followed by the variable to move and the position to move to does the same thing. Also, the command `order` followed by a list of variables will move those variable into that order and place them at the front of the variable list and `aorder` will order them all alphabetically.

Manoeuvring Between Windows

To manoeuvre between **Data editor** and **main screen**, click on the **Window** option at menu bar and from the drop down menu click on the required option (the active screen is ticked on). Alternatively choose the window from the status bar at the bottom of the screen.


Intercooled Stata 9.2 - [Results]
File Edit Prefs Data Graphics Statistics User Window Help
Results Ctrl+1
Graph
Viewer
Command Ctrl+4
Review Ctrl+5
Variables Ctrl+6
Data Editor Ctrl+7
Do-file Editor
Review
use "P:\mbbxdmg2\Fourth Year Pro
help move
edit
Copyright 1984-2006
StataCorp
4905 Lakeway Drive
College Station, Texas 77845 USA
800-STATA-PC http://www.stata.com
979-696-4600 stata@stata.com
979-696-4601 (fax)
15-user Stata for Windows (network) perpetual license:
Serial number: 1990521361
Licensed to: Epidemiology and Health Sciences
The University of Manchester
Notes:
1. (/m# option or -set memory-) 10.00 MB allocated to data
2. New update available; type -update all-
. use "P:\mbbxdmg2\Fourth Year Project Option\foundry.dta"
. help move
. edit
- preserve
Variables
idno
group
age
dtbirth
dtassmnt
dtemplmt
sex
ht
fevmeas
fevpred
fvcmeas
fvcpred
asthma
bron
smknow
smkever
empyrs
cigno
cigyrs
respdust
Command
M:\Stata9

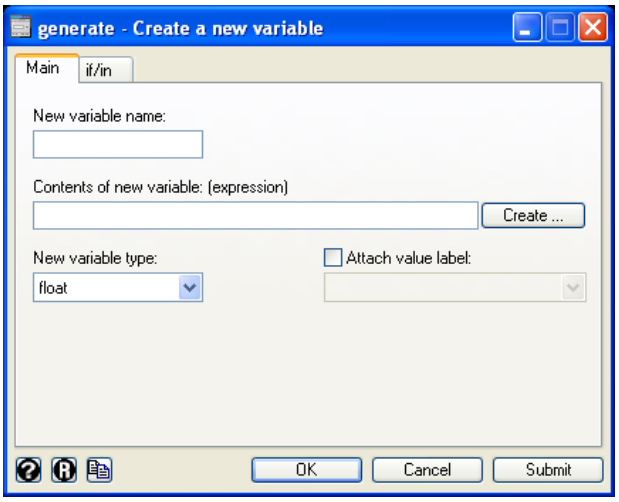
PART II

CONSTRUCTING NEW VARIABLES

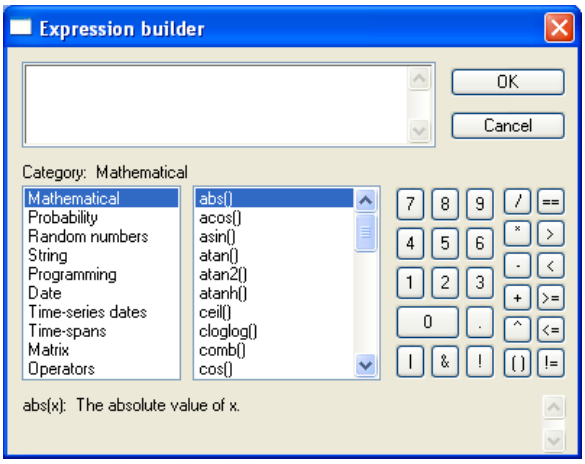
Sometimes we need to compute new variables from the data entered. For example in the foundry data set we might want to compute the ratio of the measured to predicted FEV. Alternatively we might want to group ages into bands. STATA has procedures to construct a new variable from existing variables.

Computing a New Variable

For the foundry worker data we shall compute the variable **fevratio** defined as **fevmeas/fevpred.** Click **Data, Create or change variables** then **Create new variable** and the following screen appears:-



Enter the name **fevratio** in **New variable name** box. To build up mathematical expression which will create the contents of the new variable you click the Create button. Here you can create a wide variety of expressions using the current set of variables or any of the keys on the calculator pad in the centre or any of the functions from the built-in functions box followed by.

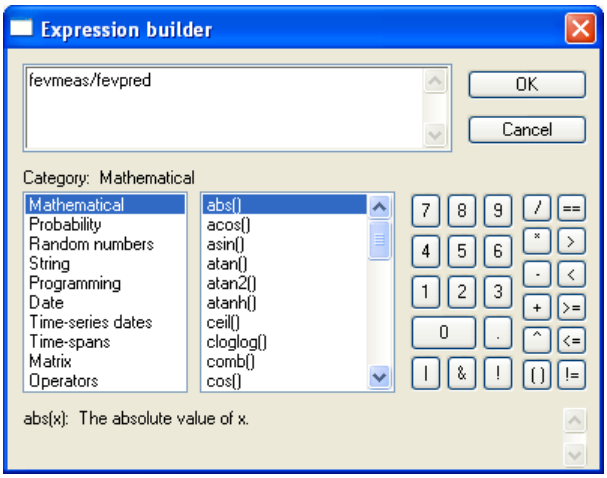


Select the function by first clicking on the category then the appropriate command, this should then appear in the **Numeric Expression** window. If a current variable is require the variable name needs to manually written into the window at the appropriate location.

These are the functions on the calculator pad are defined as follows.

Operator	Mnemonic form	Description	Operator	Mnemonic form	Description
+		Addition	>=	GE	Greater Than Or Equal To
-		Subtraction	==	EQ	Equals
*		Multiplication	!=	NE	Not Equals
/		Division	&	AND	Logical And
^		Power Of	|	OR	Logical Or
<	LT	Less Than	( )		Parentheses
>	GT	Greater Than	~	NOT	Logical Not
<=	LE	Less Than Or Equal To			

To compute **fevratio** we write **fevmeas** divided by **fevpred** into the **numeric expression** window. This is illustrated below.



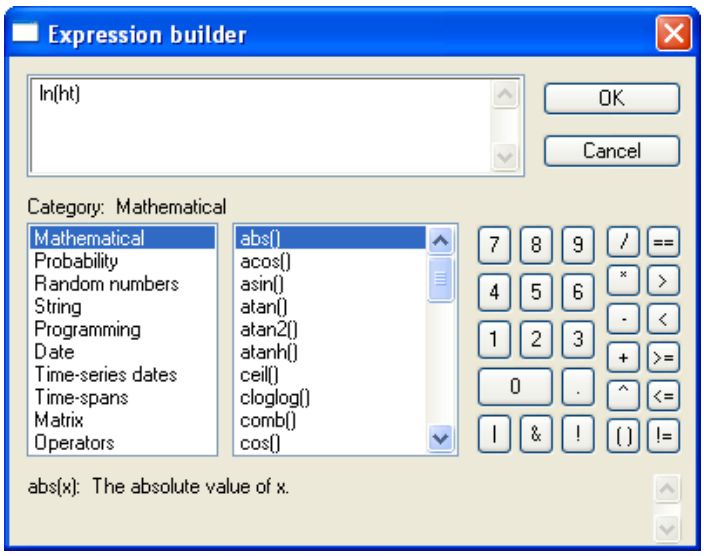
Once the expression is complete press **OK**, this will return you to the original window press **Ok** again and the new variable is generated. Any variable or value labels can be added in the same way as previously described.

The command `generate` can also be used to create a new variable, the above procedure can be performed with the following command;

```
generate fevratio = fevmeas/fevpred
```

Computing a New Variable by using built-in Functions

In the **Generate** procedure there are built in functions which can be used to create a new variable or to transform the values of an existing variable. Transformations such as the square root, or the logarithm, are easily made. Suppose you wish to do a log transformation of the variable called height (**ht**) from the **foundry** data set. Open the Expression builder using the same procedure as before, making sure to insert the new variable name as **lht.**



Click on the **Mathematical** category to scroll up and down through the mathematical functions. Select the **ln()** function for natural log and double click to put the function with a **?** in parentheses in the **Expression window.** Then select the variable to replace **?** i.e. by writing in the variable name **ht** and then press **OK** button. Then a new variable **lht** will be created (located at the end of the variable list). Having carried out a transformation it is important to check the result. For example, taking a log of a negative value creates a missing value. Other commonly used transformation functions are **lg10, sqrt, abs,** etc.

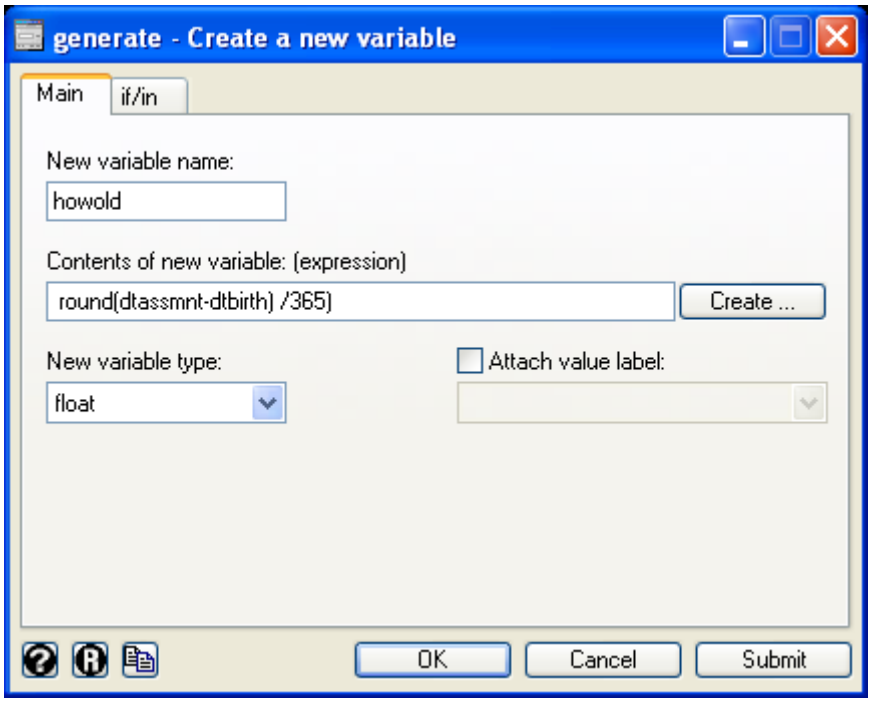
All of the functions found in the Expression window can be written into a command. Unfortunately, there are far too many to show here, however you can see them by typing the command `help function` followed by selecting the suitable category. The same command generate is employed here, the only difference is the expression will use the function.

```
generate lht = ln(ht)
```

Computing Duration of Time Difference by built-in Functions

In the same data set there are some variables (date of birth, date of assessment etc) which are stored in date format. One is able to calculate the time difference (in days) due to the way that STATA codes dates. Each date is given a number that corresponds to a specific date with 0 set to 01/01/1960 (mm/dd/yy) all dates after are given a positive number and all dates before are given a negative number. It is then easy to calculate the time elapsed (in days) by subtracting one date away from another.

The age of the patients on the date of assessment can be calculated from the date of birth and assessment date. As before click **Data>Create or change variables>Create new variable** from menu bar. After typing **howold** into the variable name box click the **Create** button to get the **expression window**. By just subtracting dtbirth away from dtassmnt we get the persons age in days, to get age in years divide by 365 and use the round function. This will give the persons age in years in integer form. Below is the example.



Whenever you compute a new variable from existing data it is important to check that what you have created is sensible. You also need to check that missing values have not been converted into none missing values. Using the **Data Browser** to check the value of **howold**.

The command is just the same as in the previous section with the exception of the use of the date command and that the command requires the variables to be string

```
generate howold = round(( dtassmnt - dtbirth)/365)
```

<u>Exercise</u> Calculate the duration of the patients in the employment and compare with the values of employment years provided in the data set.

Recoding a value

To assist in data analyses you often need to group a continuous variable (e.g. age) into categories To do this you will need to **Recode** the data**.** There are two possible options for recoding;

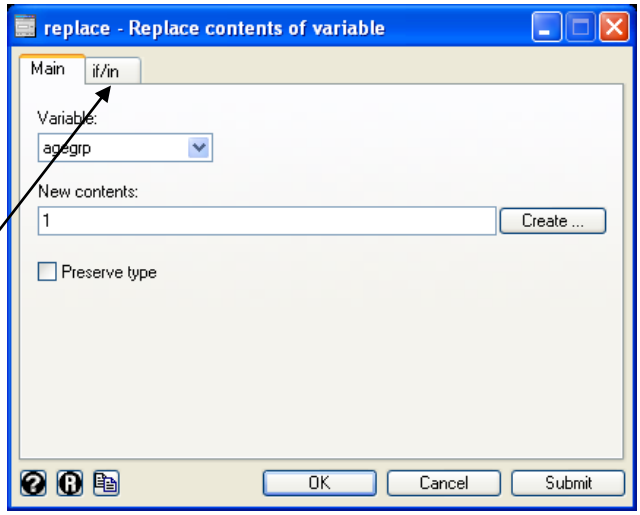
- Into Same Variables
- Into Different Variables

The first option leads to potentially valuable information being overwritten. It is usually best to use the second option as it is then possible to check whether the recode has worked correctly by comparing the new and old version.

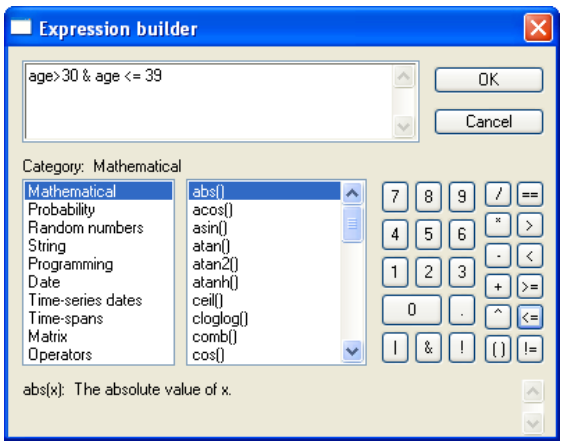
The second option requires a variable to be generated first, as shown in the previous few chapters. Suppose we wish to recode age into bands <30, 30-39, 40-49, 50+ if we generate the variable so that all the values are equal to 0 then we can assume 0 represents the first group <30 and then recode those case that fall in different age groups. To save time use the command

```
generate agegrp = 0
```

To recode click **Data** > **Create or change variables** > **Change contents of variable** to get the following screen, insert the variable to recode followed by the new contents, in this case 1 is going to equal 30-39.



To make sure that only those aged 30-39 are changed to one we use an if constraint. Click the tab **if/in** followed by **Create** and use the >, <= and & commands to specify the range 30-39, as shown below;



Now press click **Ok** through to perform the recode. As we wish to recode age into bands <30, 30-39, 40-49, 50+, repeat the process each time altering the if constraint and the change in value. This can be very repetitive and hence the `replace` command is used to speed things up.

```
generate agegrp = 0
replace agegrp = 1 if age>=30 & age <= 39
replace agegrp = 2 if age>=40 & age <= 49
replace agegrp = 3 if age>=50
```

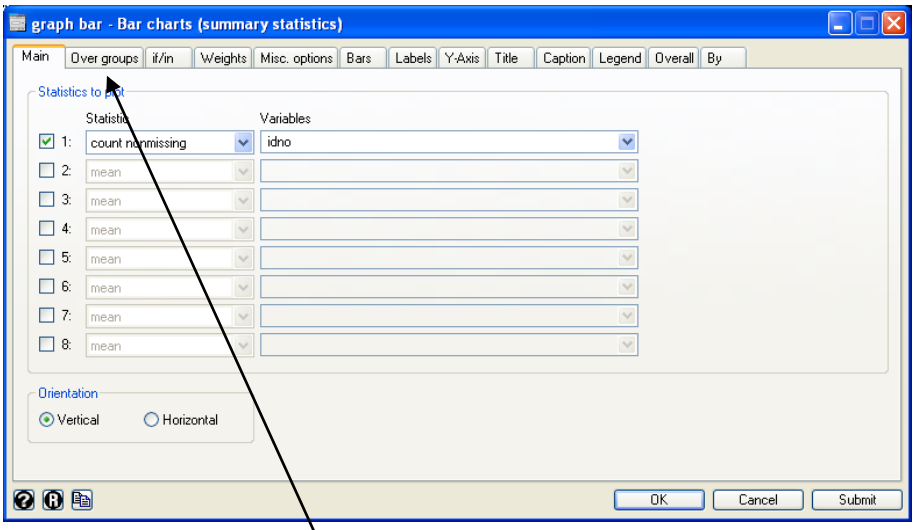
After recoding a variable you should give the numbers label using the label value command as shown previously, it is then advisable to run case summaries to compare the old and new values

GRAPHS

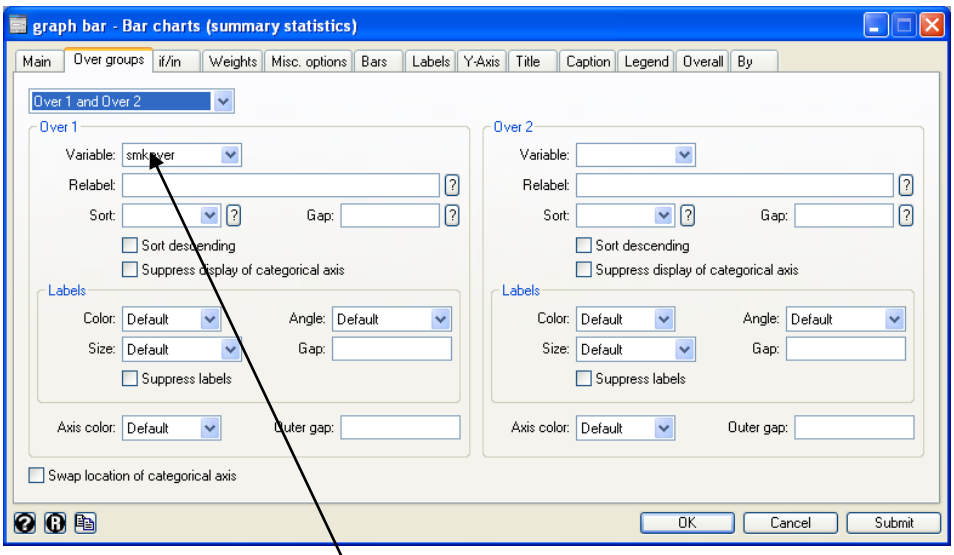
STATA will produce good quality high- resolution statistical graphics. We will look at Bar Charts, Histograms, and Scatter Plots with regression lines. In STATA graphics is the one situation where it is often better to produce them using the menus only. The commands for graphics can become very long and complicated especially when trying to alter the graphs presentation, hence it is often simpler to use the menus. In any case, the commands shall be given for all procedures in this section and it is left to the user to decide which method to take.

Bar Charts

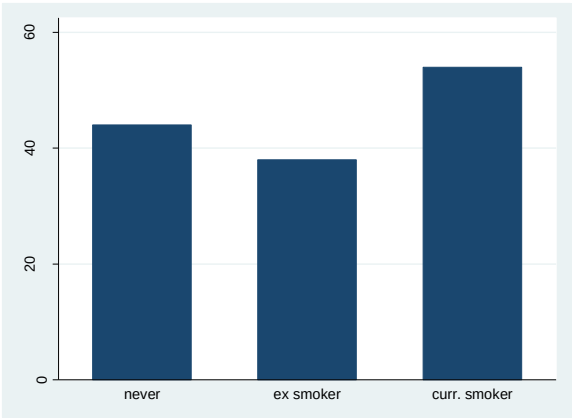
Bar Charts can only be produced for categorical variables e.g. Ever smoked Asthma etc. To produce a Bar Chart click **Graphics** > **Bar Charts** > **Summary statistics** and the following screen appears.



Alter the Statistic to **count nonmissing** and the variable to the unique identifier (**idno**), then switch to the Over groups screen by clicking the tab at the top of the window.



Here, insert the categorical variable (**smkever**) into the Variable for Over 1, this will indicate the how the bars are formed. The following simple bar graph is formed.



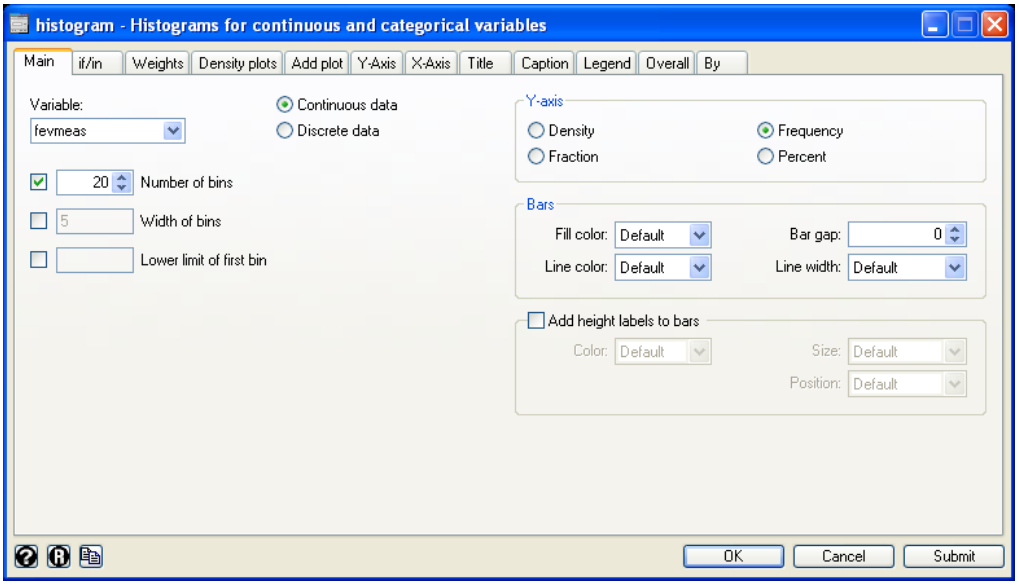
It should be noted that this bar graph is in its simplest version and that through the many tabs on the graph bar window shown on the previous page, it is possible to alter many aspects especially in terms of presentation.

The commands for all graphics within STATA begin with `graph`, which is then followed by a second command to identify which type, in this case `bar`. The command below produces the above graph.

```
graph bar (count) idno, over(smkever)
```

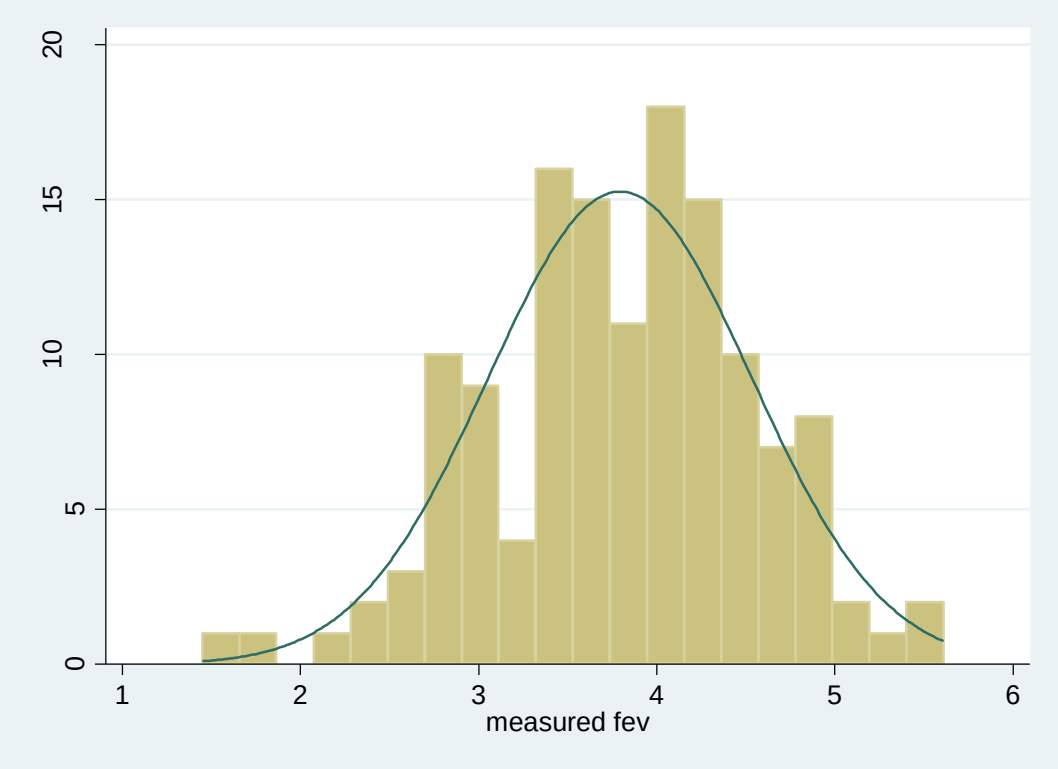
Histograms

Histograms are produced for interval variables e.g. age. To produce a histogram click on **Graphics** then **Histogram** and the following screen appears.



Click on the required variable, in this case FEV then click the on the number of bins alter to a suitable amount in this case 20 and make sure that the Y-axis is set to Frequency, finish by pressing **OK.** If you require a normal curve to be drawn on to the graph click on **Density plots** and click the option to add a **normal curve.**

This is the Histogram produced for measured FEV.



As before, the graph can be formatted to the users liking, including titles for X and Y axis and the main title

The command for this particular graph is;

```
histogram fevmeas, bin(20) frequency normal
```

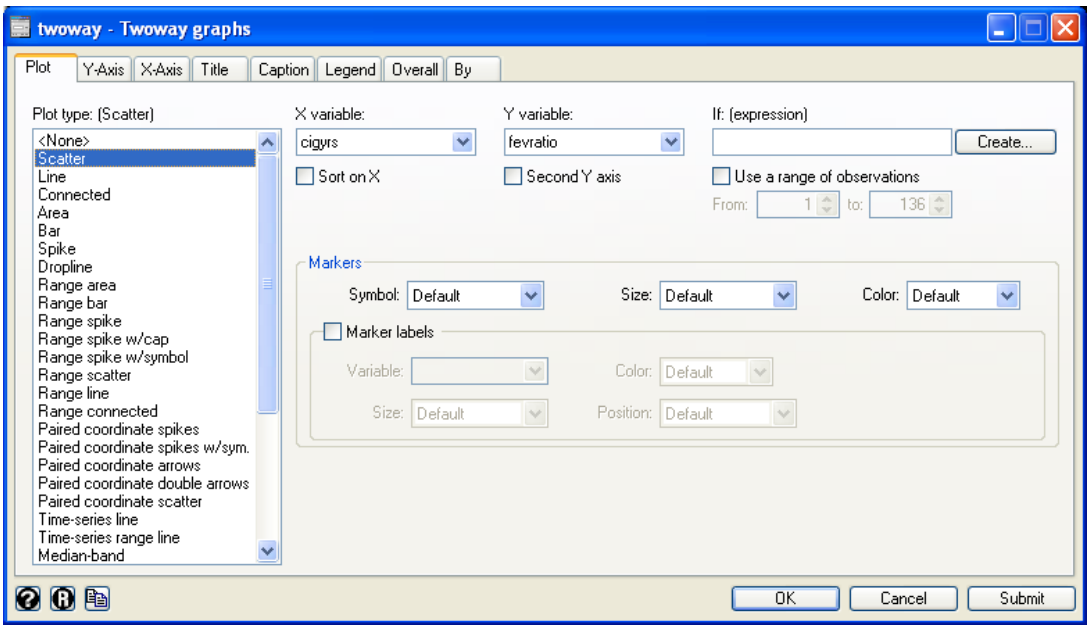
Scatter Plots

Scatter plots show the joint behaviour of two interval variables. If you want to decide whether two interval variables are related in any way you should first draw a scatter plot.

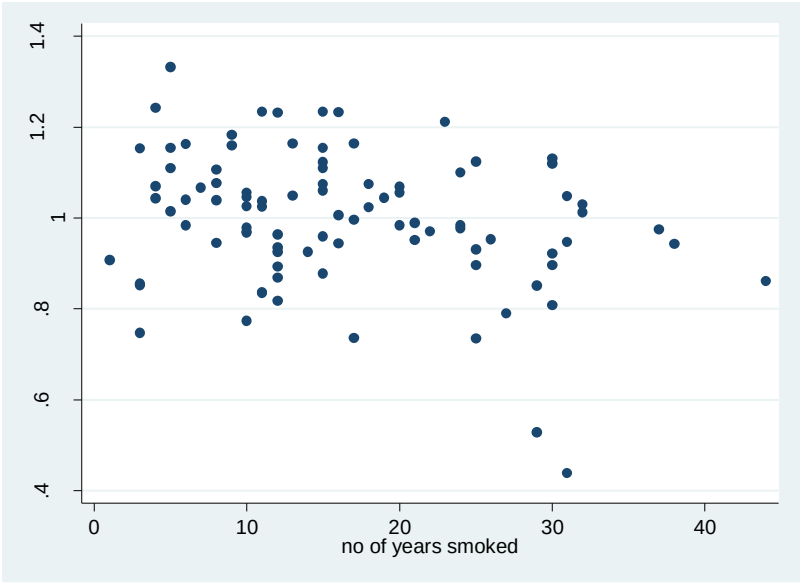
Scatter plots have 2 axes:

- the value of the dependent or response variable on the vertical y axis.
- the value of the independent variable on the horizontal x axis.

To run a scatter plot click **Graphics > Twoway Graph (Scatter, Line, etc)** and the following screen will appear. There is a list of possible graphs down the left hand side, click on **Scatter** and then select variables. In this example we choose FEV ratio as the dependent variable and the number of years smoked as the independent variable.



The above selection produces the following graph

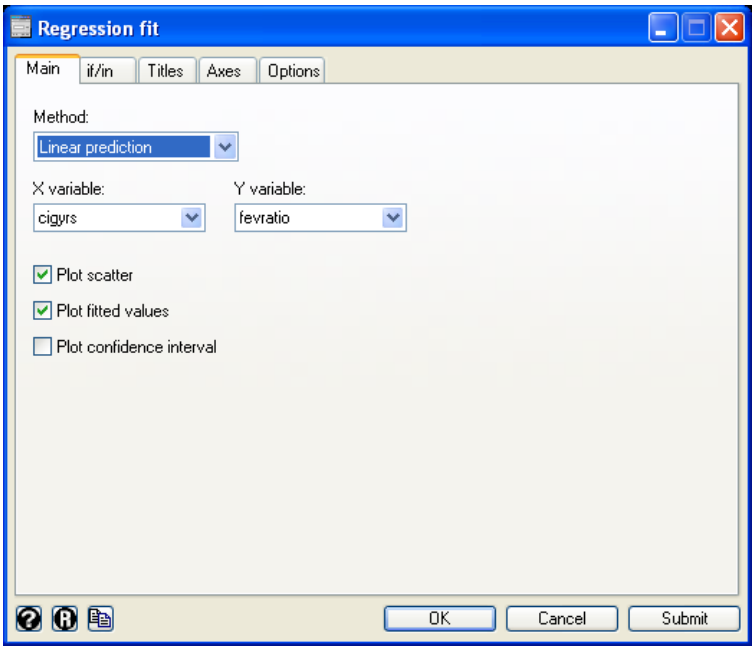


This graph is produced by the following STATA command,

```
twoway (scatter fevratio cigyrs)
```

Plotting a Regression Line on a Scatter Plot

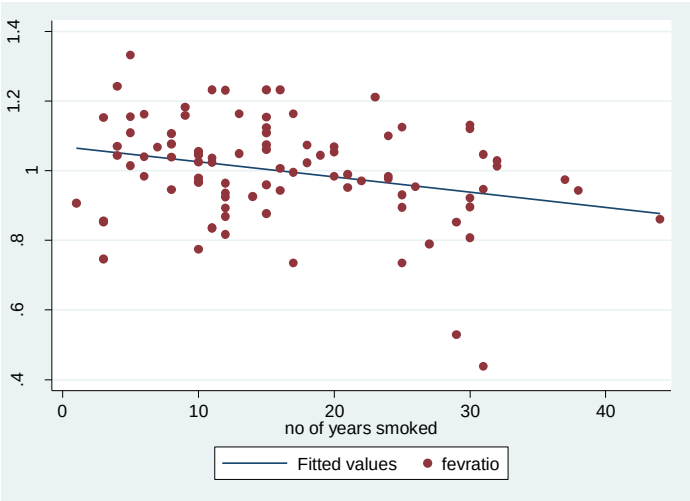
To fit a line of regression on a scatter graph an alternative menu is required, click **Graphics** > **Easy Graph** > **Regression plot**, the following screen then appears. Insert the variables and click plot scatter & Plot fitted lines.



This along with the command

twoway (lfit fevratio cigyrs) (scatter fevratio cigyrs)

produces the following graph.



This is about the limit of the STATA menu driven graphs, however if you use the commands it is possible to take this a little bit further. For example, if a plot with regression lines of FEV ratio against the number of years smoked split by exposure group was required. This can be thought of in stages.

A straight forward scatter plot of the data uses the command;

```
twoway (scatter fevratio cigyrs)
```

Using an `if` statements a plot can show just those cases who are in the exposed group

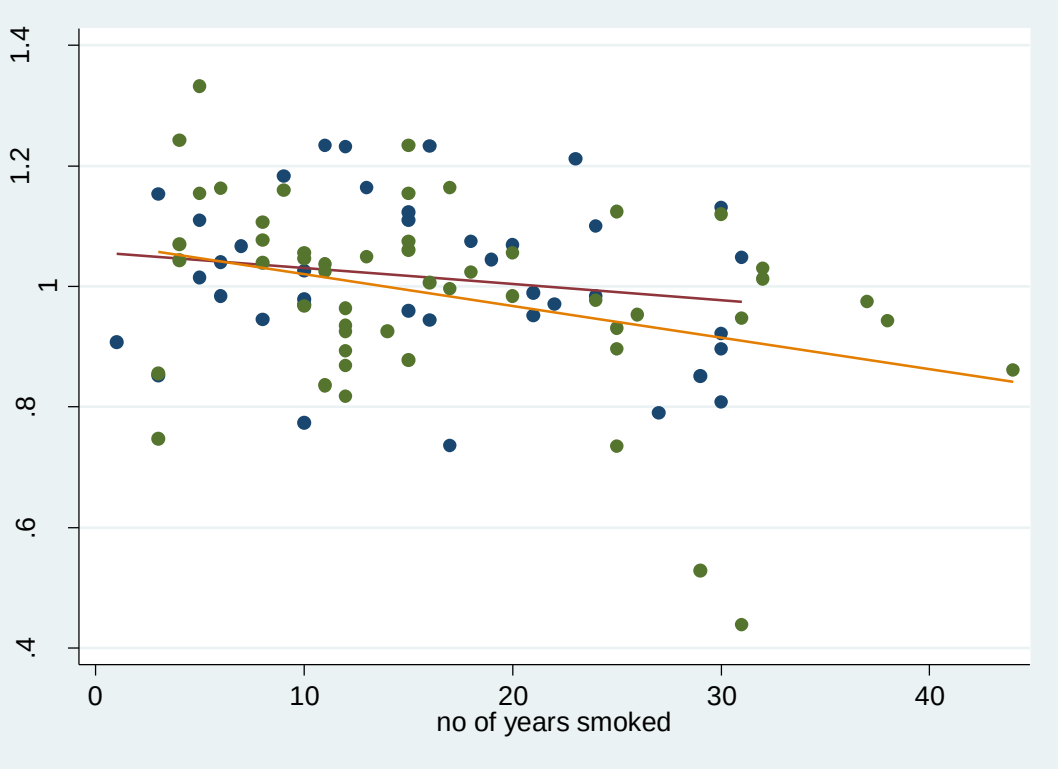
```
twoway (scatter fevratio cigyrs if group==1 )
```

where `group==1` referes to the exposed cases. If the section surrounded in brackets is repeated in the command with `group==0` then a plot that shows both groups separately can be seen

```
twoway (scatter fevratio cigyrs if group==0 ) (scatter
       fevratio cigyrs if group==1 )
```

In order to add a regression line we use the command lfit, which as with the scatter can be constrained by group, giving the command and plot below. (note `legend(off)` removes the key)

```
twoway (scatter fevratio cigyrs if group==0 ) (lfit fevratio
        cigyrs if group==0) (scatter fevratio cigyrs if
        group==1 ) (lfit fevratio cigyrs if group==1),
        legend(off)
```



STATISTICAL INFERENCE IN STATA

Introduction

This part will introduce the basic methods of statistical inference available in STATA. It will assume some familiarity with concepts in statistical inference including hypothesis testing and confidence intervals. If you are unfamiliar with these concepts, you are strongly recommended to read an introductory text in medical statistics such as Campbell and Machin “Medical Statistics A Common Sense Approach”. Some example are given at the Medical Statistics support web site at http://www.teaching-biomed.man.ac.uk/resources/informatics/statistics/

The methods will be illustrated by the Foundry data set that was considered in Part I. The purpose of this study was to examine whether dust increased respiratory morbidity. In this study the measure of respiratory morbidity are “Ever had asthma", “Ever had bronchitis”, “Measured FEV” and “Measured FVC”. The variable “Predicted FEV” and “Predicted FVC” are the values that are expected for a person’s demographic characteristics including Age, Height and Sex. Exposure to dust is measured by two variables “Exposed/Un-exposed” and dust levels recorded only for exposed workers. Because smoking is a confounding factor in this study, smoking behaviour has been recorded in terms of current smoking status (smknow), smoking history (smkever), and consumption (cigno) and duration of smoking (cigyrs).

During this part of the practical you may need to refer to the notes from Part I. If you are starting the tutorial at this point rather than continuing from Part I, you will need to open the dataset at **Shared Data > mhs > Health Methodology Course Data.**

Categorical Variable

In the first part of the study we examined whether there was any relationship between exposure to dust and smoking. Using the cross-tabs procedure we can generate the following table.

```
tabulate smknow  group, col

    do you |    exposure group
 smoke now | unexposed  exposure  |     Total
-----------+----------------------+----------
        no |        43         39 |        82
           |     68.25      53.42 |     60.29
-----------+----------------------+----------
       yes |        20         34 |        54
           |     31.75      46.58 |     39.71
-----------+----------------------+----------
     Total |        63         73 |       136
           |    100.00     100.00 |    100.00
```

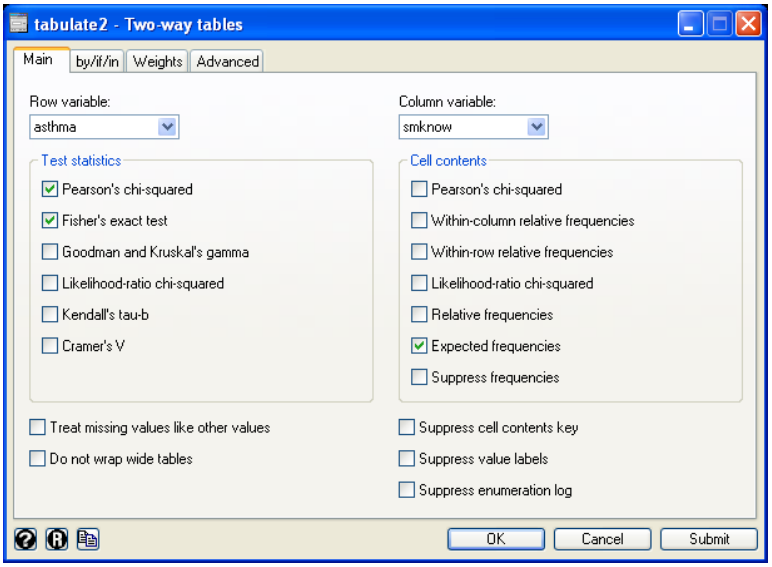
From the table above it can be seen that the percentage of workers who currently smoke is higher for those exposed to dust than those who are not, 47% as compared to 32%.

We will now examine whether respiratory symptoms as measured by the variable **asthma** relate to smoking. Using cross-tabs procedure again we obtain the following table.

```
tabulate asthma smknow, col

  ever had |   do you smoke now
    asthma |        no        yes |     Total
-----------+----------------------+----------
        no |        77         48 |       125
           |     93.90      88.89 |     91.91
-----------+----------------------+----------
       yes |         5          6 |        11
           |      6.10      11.11 |      8.09
-----------+----------------------+----------
     Total |        82         54 |       136
           |    100.00     100.00 |    100.00
```

The Chi-squared test and Fisher's Exact test

Amongst those who currently smoked 11.1% had experienced symptoms of asthma whilst only 6.1% amongst those who did not. Does this suggest that smoking may be related to asthma or might this difference be due to chance - that is explained by sampling variation? One way in which we can examine this is by a chi-squared test. This can be carried out by re-running the cross-tab procedure including the chi-squared statistics option as follows. In the cross-tabs panel (**Statistics > Summaries, Tables & Tests > Tables > Twoway tables with measures of association**) we select under Test Statistics **Pearson's chi-squared** and **Fisher's exact test**. Under Cell contents select **Expected Frequencys.**



Then click on **OK** to get the analysis below

```
+-------------------+
| Key               |
|-------------------|
|     frequency     |
| expected frequency |
+-------------------+

  ever had |   do you smoke now
    asthma |        no        yes |     Total
-----------+----------------------+----------
        no |        77         48 |       125
           |      75.4       49.6 |     125.0
-----------+----------------------+----------
       yes |         5          6 |        11
           |       6.6        4.4 |      11.0
-----------+----------------------+----------
     Total |        82         54 |       136
           |      82.0       54.0 |     136.0

          Pearson chi2(1) =   1.1009   Pr = 0.294
           Fisher's exact =                 0.344
   1-sided Fisher's exact =                 0.231
```

The panel above gives the results of a chi-squared test of no association between asthma and smoking. In interpreting this table, one is concerned with the three probabilities shown in the bottom right corner. These are the p-values for the significance test. Firstly it is usually recommended that you consider a 2-sided rather than 1-sided test. As one of the cells has an expected count less than or equal to 5 and it is a 2 by 2 table, it is recommended that we look at the Fisher Exact test which provides the valid result of 0.344. Assuming the conventional 0.05 significance level, this result is considered non-significant. In reporting results of statistical tests you are strongly recommended to give the p-value rather than just write “significant” or “non-significant”. In reporting this we might write “there was no evidence of an association between smoking and asthma (Fishers exact p-value=0.344)." Had the expected count been greater less than 5 and it is suggested that you report the Chi-squared test p-value.

The Stata command for a cross tabulation is used to produce this result with some of the extra options added into the command in this case `chi2 exact expected` which add a chi-square test, fishers exact test and a set of expected values on to the analysis;

```
tabulate asthma smkever, chi2 exact expected
```

Exercise Using the cross-tabs procedure examine whether there is a relationship between current smoking status and bronchitis symptoms.

Are the expected numbers greater than 5 for all cells?

Fill in the spaces and delete as appropriate in the following statement:

"Amongst those that currently smoked ___% had experienced symptoms of bronchitis whereas ___% of non-smokers experience such symptoms. This was statistically significant/non significant at a 5% level using a two-tailed continuity corrected chi-squared test with p=_____ "

Exercise Now use the cross-tabs procedure to examine the relationship between Exposure to dust and symptoms of bronchitis and asthma. Record your conclusions below using either the continuity corrected chi-squared or Fisher's exact test as appropriate.

We have found no statistically significant relationship between exposure to dust and either asthma or bronchitis symptoms. For bronchitis symptoms you should have obtained the following tables using the command;

```
tabulate bron group, chi2 exact expected.
```

```
+--------------------+
| Key                |
|--------------------|
|     frequency      |
| expected frequency |
+--------------------+

  ever had |    exposure group
bronchitis | unexposed  exposure  |     Total
-----------+----------------------+----------
        no |        59         62 |       121
           |      56.1       64.9 |     121.0
-----------+----------------------+----------
       yes |         4         11 |        15
           |       6.9        8.1 |      15.0
-----------+----------------------+----------
     Total |        63         73 |       136
           |      63.0       73.0 |     136.0

          Pearson chi2(1) =   2.6199   Pr = 0.106
           Fisher's exact =                 0.169
   1-sided Fisher's exact =                 0.088
```

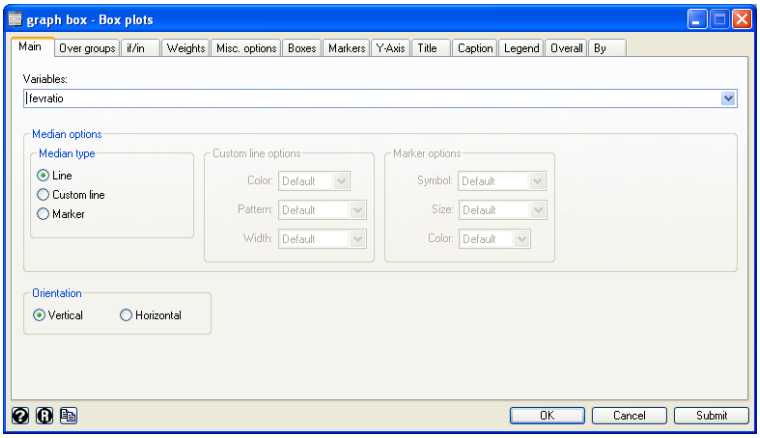
Whilst 15% (11/73) of the exposed worker had symptoms of bronchitis and only 6% (4/63) of non-exposed, this difference was not statistically significant at the 5% level (Fishers Exact test p=0.169, due to 2x2 and no continuity correction). There are several explanations for this. There may be no relationship between the exposure to dust and respiratory disease. Alternatively, the study may have lacked statistical power to detect small differences. It should be noted also that only 11% (15/136) of the sample reported such symptoms.

CONTINUOUS OUTCOME MEASURES

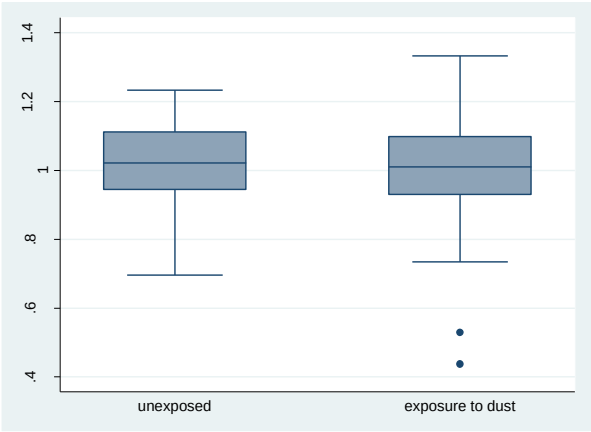
We will now consider the lung function measurements. Given that lung function is age and size dependent it is usual to divide measured lung function by the expected lung function. In Part I we constructed such a variable.

Exercise Using the Generate command construct new variable **fevratio** and **fvcratio** defined by **fevmeas/fevpred** and **fvcmeas/fvcpred**.

We now want to examine whether workers exposed to dust have reduced lung function. First we might examine this graphically with a box plot. Going to the graphics menu, select **boxplot** and fill in with the main variable as **fevratio** and the Over groups variable (second tab) as groups;



Along with the command - `graph box fevratio, medtype(line) over(group)` gives the following plot;



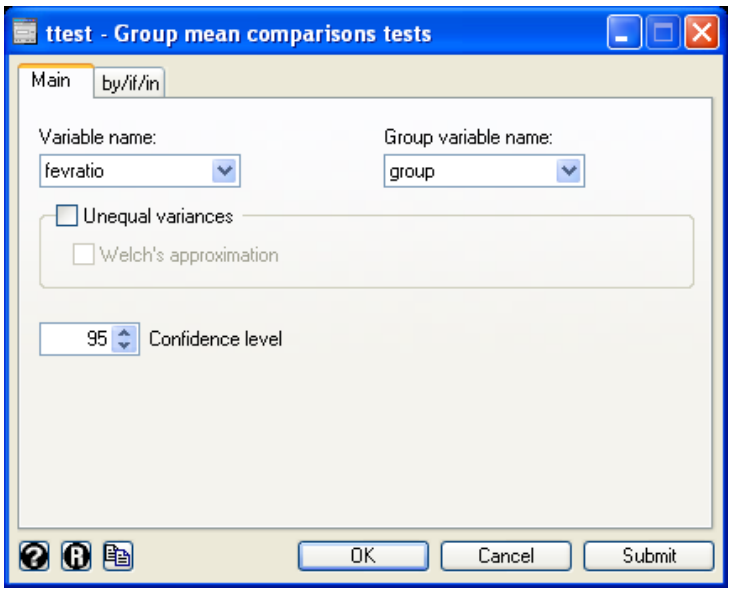
The box represents the inter-quartile range; the whiskers represent the range. The solid line in the middle represents the median. This suggests that there is little difference between the dust exposed and non-exposed workers. Other **Analysis** options we might use to compare the lung function of exposed and non-exposed workers are **Summary statistics** in the **summarise, tables & tests section**.

Exercise Use **summarize** and **by** options to compare lung function of exposed with non-exposed workers using fvcratio and fevratio. Record the results below.

	Mean	*Standard Deviation*	*Median*	*Max*	*Min*	*N*
Exposed						
Non Exposed						

Comparison of Means Using a t-test

The t-test procedure can be used for statistical comparison of the mean **FEV ratio** of the exposed compared to non-exposed workers. It will also give the confidence interval for the difference of the two means. For the test go to **Statistics** > **Summary, tables & tests** > **Classic tests of hypothesis** > **Group mean comparison** the following panel then appears into which we have selected **fevrat** as the variable and group as the group variable name.



Clicking **Ok** gives the results below. Note, first you will need to test the assumption of equal variance using a Levene's test, this is done in exactly the same way as above except you choose **Group variance comparison test** instead of Group means comparison test and fill in as above.

The command for the t-test and variance test is `ttest` and `sdtest` respectively, in order to perform the analysis on a two predefined groups we use the `by` command. Therefore the corresponding command for the above procedure is;

```
sdtest fevratio, by(group)
ttest fevratio, by(group)
```

The first table gives the Levene's F-Test of equality of variance – the assumption of a t-test is that the means for each group have the same variance. For this data there is no evidence that the variance as p=0.2413 for the Levene's test.

```
Variance ratio test
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
unexpose |      63    1.015766    .0161072    .1278466    .9835679    1.047963
exposure |      73    1.000312     .017309    .1478885    .9658066    1.034816
---------+--------------------------------------------------------------------
combined |     136     1.00747    .0118913    .1386754    .9839531    1.030988
------------------------------------------------------------------------------
    ratio = sd(unexpose) / sd(exposure)                           f =   0.7473
Ho: ratio = 1                                     degrees of freedom =   62, 72
    Ha: ratio < 1               Ha: ratio != 1                 Ha: ratio > 1
  Pr(F < f) = 0.1206         2*Pr(F < f) = 0.2413           Pr(F > f) = 0.8794
```

The second table gives a t-test for equal, note if unequal variances are found the test is easily altered by clicking the unequal button on the window or typing `unequal` as an option to the command. The t-test results although in this case it makes little difference. The result can be summarised as "there was no evidence of increased FEV ratio for workers exposed to dust (mean diff=0.0155, 95% c.i -0.032 to 0.063 p=0.519)"

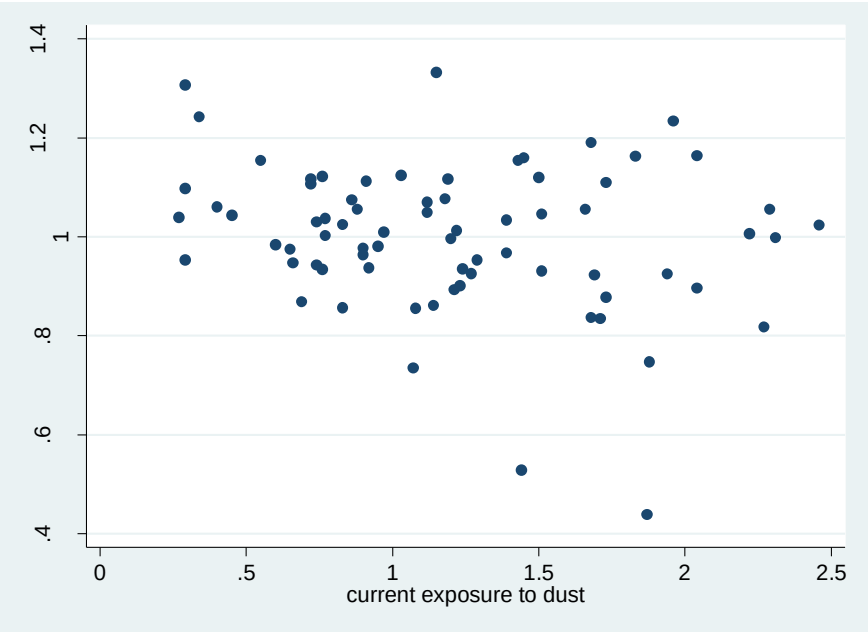
```
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
unexpose |      63    1.015766    .0161072    .1278466    .9835679    1.047963
exposure |      73    1.000312     .017309    .1478885    .9658066    1.034816
---------+--------------------------------------------------------------------
combined |     136     1.00747    .0118913    .1386754    .9839531    1.030988
---------+--------------------------------------------------------------------
    diff |            .0154541    .0238987               -.0318134    .0627217
------------------------------------------------------------------------------
    diff = mean(unexpose) - mean(exposure)                        t =   0.6467
Ho: diff = 0                                     degrees of freedom =      134

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.7405         Pr(|T| > |t|) = 0.5190          Pr(T > t) = 0.2595
```

<u>Exercise</u> Compare mean FVC ratio for the exposed and non-exposed subjects using a t-test

From the analyses there appears to be no evidence that exposure to dust affects respiratory function. It may be argued nevertheless that being categorised as "exposed" or "not exposed" is a crude assessment for exposure. Dust exposure has been recorded for subjects in the exposed group. We will now carry out some analysis on just the exposed subjects. This can be done in any analysis by using the if condition by setting it to group=1 any analysis with this condition will only be on the dust exposed group.

Below displays a scatter plot of FEV ratio compared to dust for subjects for the exposed group.

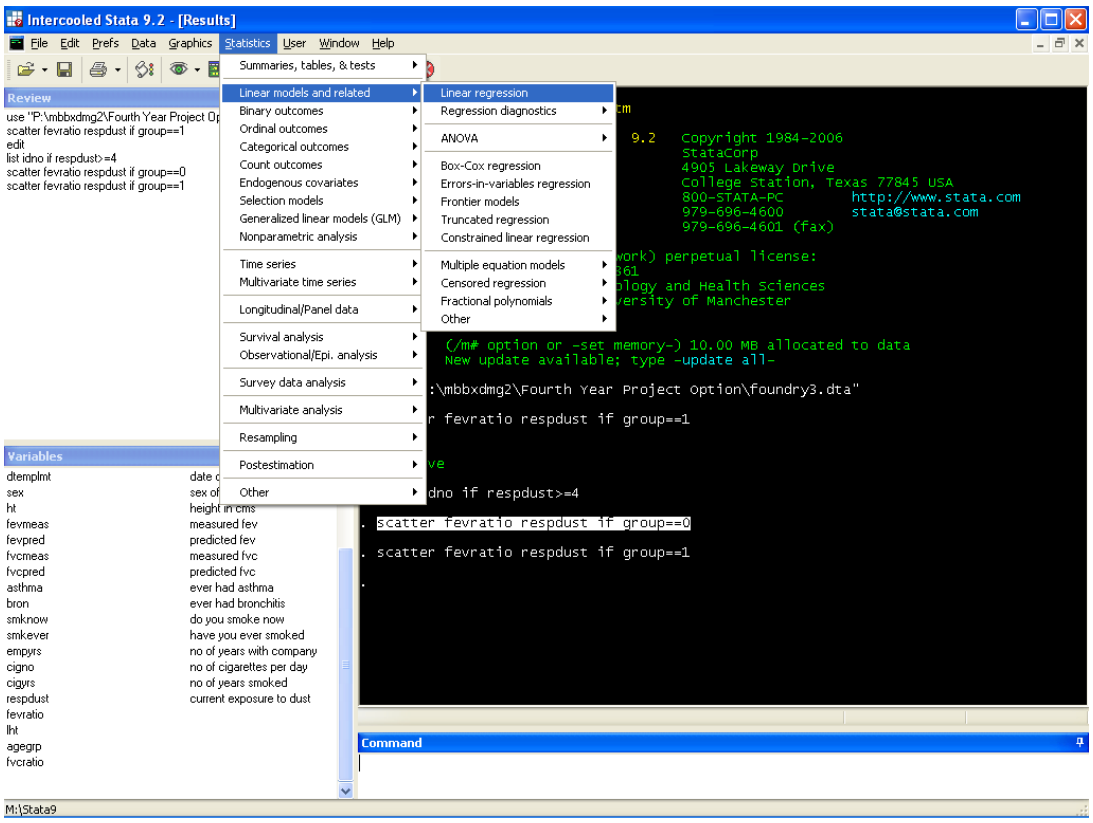
```
scatter fevratio respdust if group==1
```



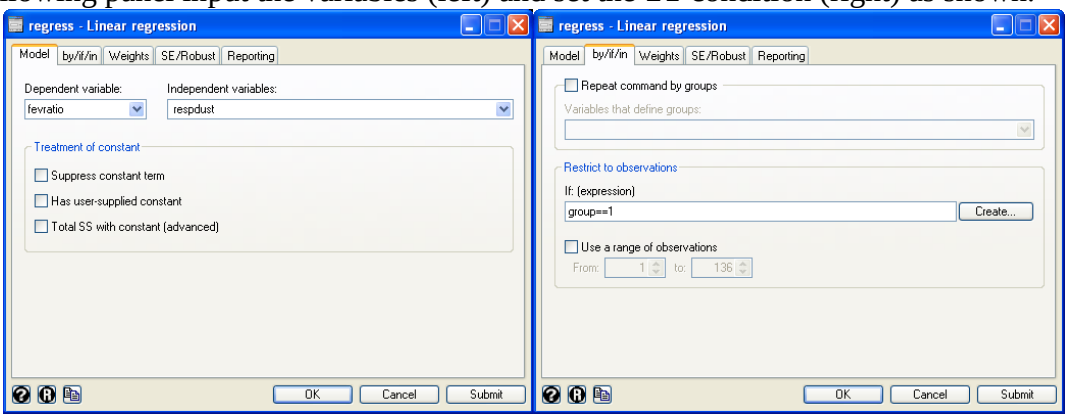
There is some suggestion from this that respiratory function may be reduced for those with higher exposure.

LINEAR REGRESSIONS

To test this we will use linear regression to fit a straight line of the form Y=A + BX, where Y is the dependent variable **fevratio** and X is independent variable **respdust**. If the gradient B is negative, this would indicate reduced respiratory function with increased dust. To do this in STATA go to the **Linear models and related** then **Linear Regression** as shown.



In the following panel input the variables (left) and set the `if` condition (right) as shown.



The following table of results is produced by the linear equation option or the regress command with the appropriate `if` statement;

```
regress fevratio respdust if group==1
```

The coefficients are the values of A and B in the equation of the line **fevratio=A+B.respdust**

```
      Source |       SS       df       MS              Number of obs =      73
-------------+------------------------------           F(  1,    71) =    3.35
       Model |  .070933055     1  .070933055           Prob > F      =  0.0714
    Residual |  1.50378031    71  .021180004           R-squared     =  0.0450
-------------+------------------------------           Adj R-squared =  0.0316
       Total |  1.57471336    72  .021871019           Root MSE      =  .14553

------------------------------------------------------------------------------
    fevratio |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    respdust |  -.0568943   .0310891    -1.83   0.071    -.1188842    .0050955
       _cons |   1.068709   .0410735    26.02   0.000     .9868113    1.150608
------------------------------------------------------------------------------
```

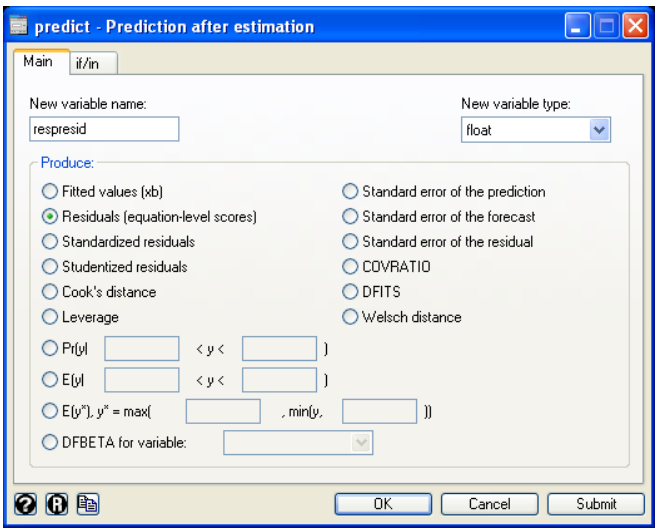
The coefficient for respiratory is equal to –0.0569 indicating a slight negative affect as current exposure to dust increases by one unit. The column labelled "`P>|t|`" gives the p-value for the statistical test that the regression coefficients differ from zero. This tells us that the constant is significantly different from zero which is not particularly interesting as we do not expect the intercept of the line with the y-axis to be zero. It also gives a p-value of 0.071for the test that the gradient differs from zero. There is some suggestion of a negative gradient, but this is not significant at the conventional 5% significance level.

The table reproduced also tells how well the line fits the data. The result for R^2 (written "R-square") is 0.045. This is an estimate of the proportion of the variance explained by the model. A line that fits the data perfectly will have an R^2 equal to 1. Where as a line that does not explain anything in the data will have an R^2 of zero. A value of R^2 equal to 0.045 is therefore not at all good – only 4.5% of the variation in the data is being explained.

The conclusion that can be drawn from this is that whilst there is a slight suggestion of reduced respiratory function with increased dust exposure, the evidence is weak.

Model Checking

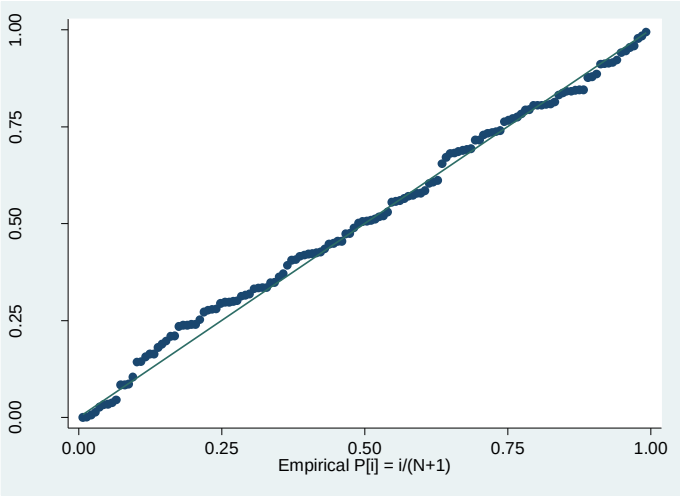
The linear regression model described by the coefficients allows one to estimate a predicted value. The difference between the observer value and the predicted value is called a residual. Where a model fits badly the regression line will have large residuals. If we consider the scatter plot above for FEV ratio compared to respiratory dust the residuals will be large. One of the assumptions of a regression model is that the residuals will have a normal distribution. One way to check this graphically is to use **normal probability plot**. This compares the residuals against a normal distribution. Such a plot can be obtained post linear regression in STATA by first creating a set of residuals by clicking **Statistics > Post estimation > Residuals, predictions, etc** to get.



Just select give a new variable name and decide what residuals are required and click **Ok**. The command

```
predict respresid, residuals
```

A normal probability plot is the created through **Statistics > Summaries, table & tests > Distributional plots & tests > Normal probability plot** insert the variable you just created to represent the residuals. The plot below is then created, if the residuals are normally distributed the plotted points are on the diagonal line. The plot below suggests that the data are approximately normally distributed. If the data were skewed the points would bulge away from the line.



<u>Exercise</u> Examine the relationship between FVC ratio and dust levels using the methods above.

NON-PARAMETRIC METHODS

Where data is not normally distributed, statistical analyses that assume a normal distribution may be inappropriate. This is especially a concern where the sample size is small (<50 in total). Variables that are discrete (take only integer values) or have an upper or lower limit are by definition non-normal. Sometimes the distribution of the data is approximately normal so this is not a problem, particularly where the sample size is large, but for some variables it may be unreasonable to treat the data as normally distributed. To illustrate this we will compare the number of cigarettes smoked by "exposed" and "non-exposed" workers who currently smoke.

You will need to select all cases that currently smoke by setting the if command to `smknow==1` as discussed previously. The frequency table for cigs per day for current smokers is given below.

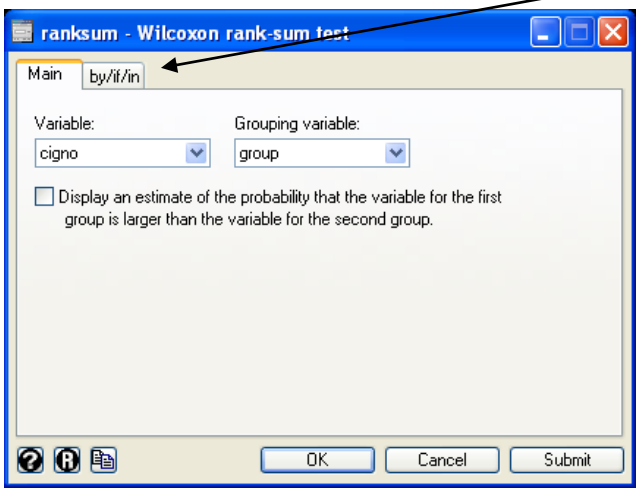
```
tab  cigno if smknow==1

      no of |
 cigarettes |
    per day |      Freq.     Percent        Cum.
------------+-----------------------------------
          3 |          2        3.70        3.70
          5 |          1        1.85        5.56
          6 |          1        1.85        7.41
         10 |          3        5.56       12.96
         12 |          2        3.70       16.67
         15 |          6       11.11       27.78
         18 |          1        1.85       29.63
         20 |         23       42.59       72.22
         25 |          6       11.11       83.33
         30 |          7       12.96       96.30
         40 |          2        3.70      100.00
------------+-----------------------------------
      Total |         54      100.00
```

More than half the sample (30/54) give values of 20 or 30 cigs. per day. Hence the variable is not even approximately normally distributed.

Exercise Use the **`summarize`** menu/command (include the `detail` option command) determine the median and inter-quartile range for **No Cigs** consumed for Exposed and Non-dust exposed workers.

Suppose we wanted to compare the median number of cigarettes smoked per day by smokers according to dust exposure group. The method one uses is the Mann-Whitney U-test, which is called a rank based **non-parametric** method. The analysis is based not on the raw data values but on the ranks of the data. The procedure ranks the values of numbers of cigarettes smoked from smallest to largest.

The Mann-Whitney U-Test is carried out as follows. Under **Statistics** select **Summaries, tables & tests** > **Non parametric tests of hypothesis** > **Man Whitney two sample ranksum test** to give a the non parametric procedure. Insert the variables as shown below remembering to add the **if** condition **`smknow==1`**.



This along with the command

```
ranksum cigno if smknow==1, by(group)
```

generates the following output

```
Two-sample Wilcoxon rank-sum (Mann-Whitney) test

       group |      obs    rank sum    expected
-------------+---------------------------------
   unexposed |       20         509         550
exposure to  |       34         976         935
-------------+---------------------------------
    combined |       54        1485        1485

unadjusted variance      3116.67
adjustment for ties      -256.25
                      ----------
adjusted variance        2860.42

Ho: cigno(group==unexposed) = cigno(group==exposure to dust)
             z =  -0.767
    Prob > |z| =   0.4433
```

In the tables above note the mean rank for each group and the significance level. The mean rank is slightly lower for the unexposed group but this is not statistically significant at a 5% significance level. Hence, we conclude that there is no difference between the median number of cigarettes smoked by "exposed" and "non-exposed" workers. The next analysis will include all subjects from the data indicating that there will be no need to use the if constraint.

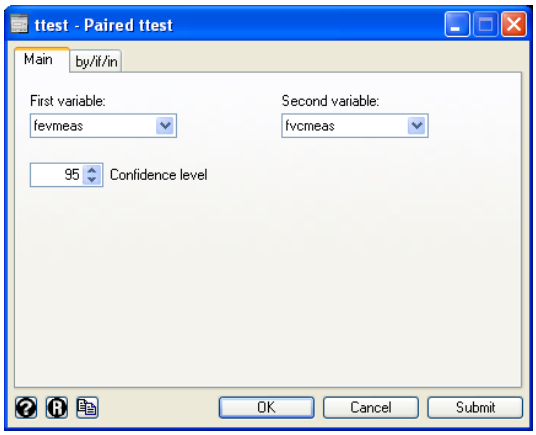
COMPARISONS OF RELATED OR PAIRED VARIABLES

For most of the analysis above we have compared the "exposed" and "non-exposed" groups of workers. In some circumstances we want to compare measures within the same subject. Such comparisons are sometimes referred to as **paired** or **pair-matched.**

Continuous Outcome Measures

One might want to compare the mean of a continuous measure at one time point with the mean of the same measure at a different time point. Whilst this may not be a sensible analysis for this data, we can illustrate this for a continuous variable by comparing FEV measured with FVC measured.

To compare the mean measured FEV with mean measured FVC we select a **Paired samples T-test** in the **Compare means** submenu. This gives the panel below. Pairs of variables are selected by highlighting the pair of variables in the window to the left then clicking on the select button to transfer to the **Paired Variable** window as shown.



Results and command are given below

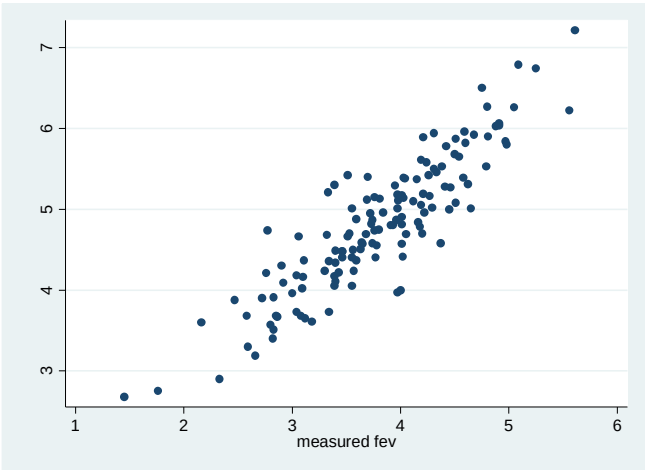
```
      ttest fevmeas == fvcmeas

Paired t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
 fevmeas |     136    3.793824       .0634    .7393642    3.668438    3.919209
 fvcmeas |     136    4.813456    .0720018    .8396775    4.671059    4.955853
---------+--------------------------------------------------------------------
    diff |     136   -1.019632    .0319845    .3730006   -1.082888   -.9563768
------------------------------------------------------------------------------
     mean(diff) = mean(fevmeas - fvcmeas)                          t = -31.8789
 Ho: mean(diff) = 0                              degrees of freedom =      135

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

It is readily apparent that mean *measured FVC* is greater than mean *measured FEV***.** We could report this as "Measured FVC was significantly higher than measured FEV (diff=1.02, 95% c.i. 0.96 to 1.08, p<0.0001)"

<u>Exercise</u> Compare the mean FEV ratio with the mean FVC ratio.

The above method of analysis compares the mean value for the two variables. It does not tell how close individual values are for the same subject. A visual way in which one can do this is with a scatter plot of the two variables as shown below. We get a visual impression that FEV and FVC are quite strongly correlated. By choosing the same numerical range for both axes we can see also that the values for FVC are systematically larger than for FEV.

```
scatter fvcmeas fevmeas
```



Analysis of Related Binary Outcomes

Suppose we wish to compare the proportion of workers who had bronchitis symptoms with the proportion who had asthma symptoms. One might first construct the cross-tabulation using the cross tabs procedure. Both row and column percentages have been added.

```
tab bron asthma, row col
```

```
  ever had |    ever had asthma
bronchitis |        no        yes |     Total
-----------+----------------------+----------
        no |       113          8 |       121
           |     93.39       6.61 |    100.00
           |     90.40      72.73 |     88.97
-----------+----------------------+----------
       yes |        12          3 |        15
           |     80.00      20.00 |    100.00
           |      9.60      27.27 |     11.03
-----------+----------------------+----------
     Total |       125         11 |       136
           |     91.91       8.09 |    100.00
           |    100.00     100.00 |    100.00
```

Careful examination of this table reveals that 11% (15/136) of workers reported bronchitis whilst only 8% (11/136) had asthma. These two proportions can be compared using McNemar's test. This is available only through the command corresponding to a Matched Case Control data (mcc) In the command select the pair of variables in the same way as for a paired t-test and as shown below.

```
mcc asthma bron
```

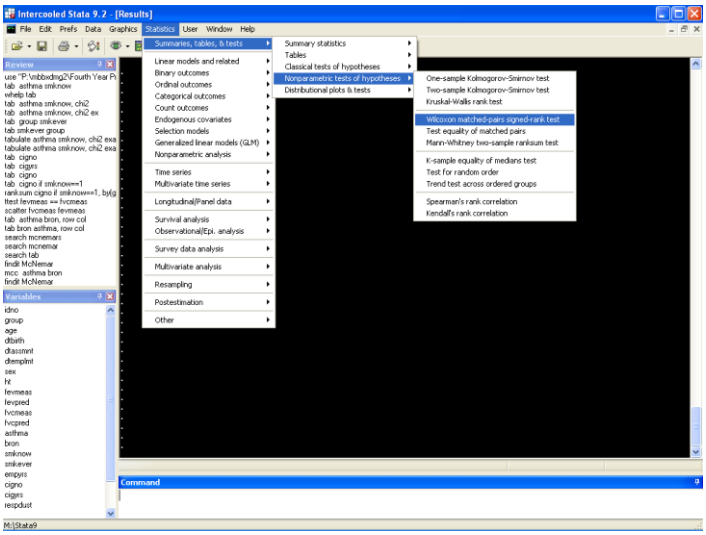
This gives the following results

```
                 | Controls                |
Cases            |   Exposed   Unexposed   |      Total
-----------------+-------------------------+------------
         Exposed |         3           8   |         11
       Unexposed |        12         113   |        125
-----------------+-------------------------+------------
           Total |        15         121   |        136

McNemar's chi2(1) =      0.80    Prob > chi2 = 0.3711
Exact McNemar significance probability       = 0.5034

Proportion with factor
        Cases       .0808824
        Controls    .1102941     [95% Conf. Interval]
                   ---------     --------------------
        difference -.0294118     -.1010251   .0422015
        ratio       .7333333       .370639   1.450948
        rel. diff. -.0330579     -.1066853   .0405696

        odds ratio  .6666667      .2363844   1.773597   (exact)
```

The p-value for the McNemar test is not significant (p=0.503) so we conclude that symptoms of bronchitis are no more common in this population than symptoms of asthma.

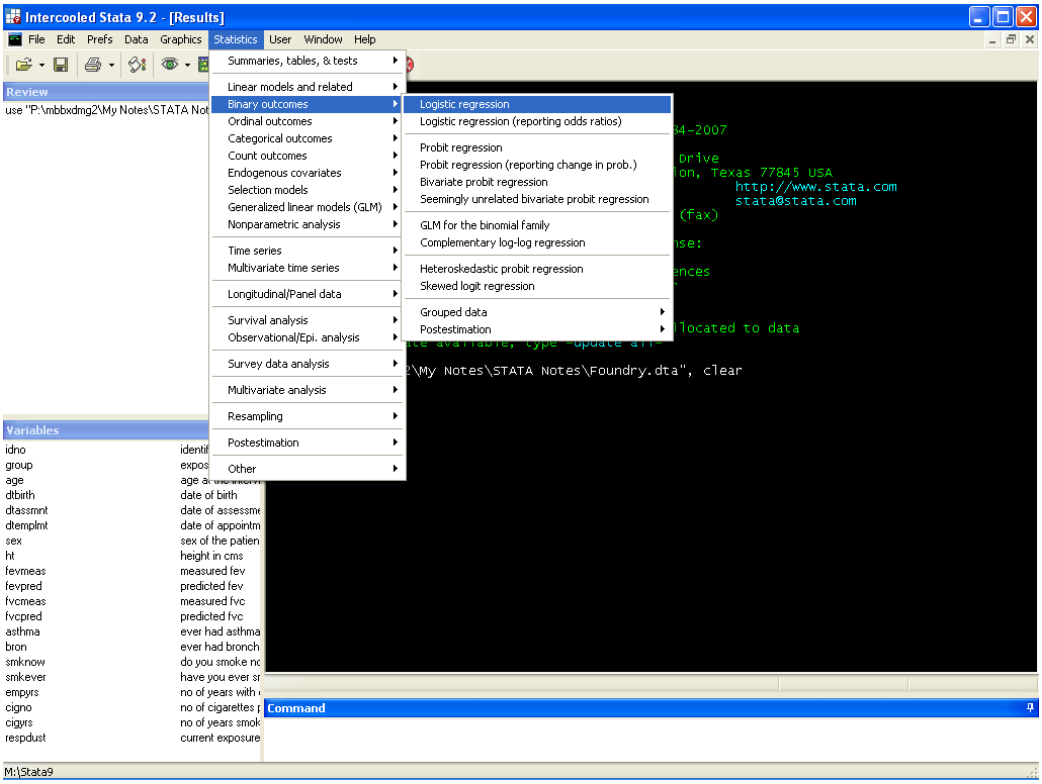
Related Ordinal Data

For ordered categorical or quantitative variables that are not plausibly normal the suggested procedure is to use the **Wilcoxon** procedure. This is selected from **Statistics > Summaries, tables & tests > Nonparametric tests of hypothesis > Wilcoxon matched pairs signed rank test** (see below).

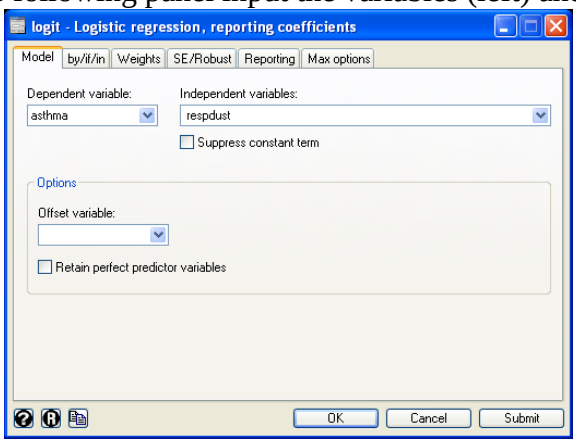


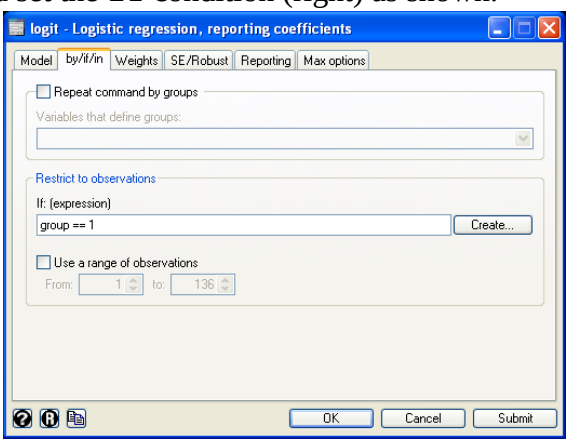
LOGISTIC REGRESSIONS

It maybe the case that we wish to model a variable which has a binary outcome, in our dataset this could be the variables regarding ever had bronchitis or asthma were the outcome is either yes or no. In this situation, the assumptions of linear regression do not apply and hence logistic regression is employed. Logistic regression uses a link function to convert the binary dependent variable into a probability, which then allows it to be fitted in a straight line form g(Y)=A + BX, where g(Y) is the probablitly of the dependent variable **asthma** being yes and X is independent variable **respdust**. If the gradient B is positive, this would indicate increased probability of asthma with increased dust. To do this in STATA go to the **Linear models and related** then **Linear Regression** as shown.



In the following panel input the variables (left) and set the `if` condition (right) as shown.





The following table of results is produced by the linear equation option or the `logit` command with the appropriate `if` statement;

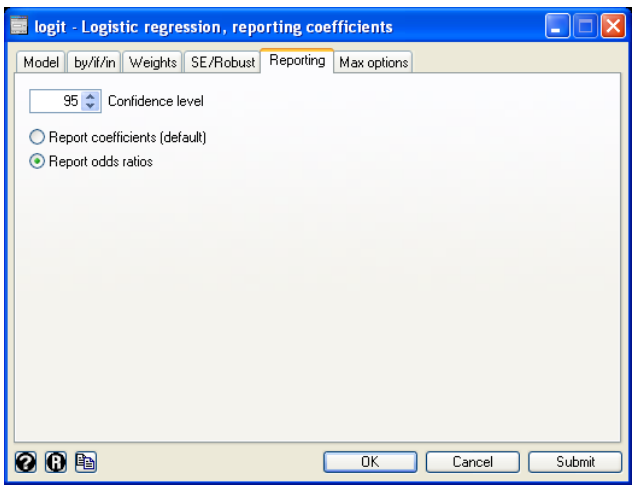
```
logit asthma respdust if group == 1
```

The coefficients are the values of A and B in the equation of the line **asthma=A+B.respdust**

```
Iteration 0:   log likelihood = -23.064955
Iteration 1:   log likelihood = -19.864628
Iteration 2:   log likelihood = -19.134899
Iteration 3:   log likelihood = -19.118417
Iteration 4:   log likelihood =  -19.11838

Logistic regression                               Number of obs   =         73
                                                  LR chi2(1)      =       7.89
                                                  Prob > chi2     =     0.0050
Log likelihood =  -19.11838                       Pseudo R2       =     0.1711

------------------------------------------------------------------------------
      asthma |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    respdust |   2.151492   .8514392     2.53   0.012     .4827015    3.820282
       _cons |  -5.370413   1.484742    -3.62   0.000    -8.280455   -2.460371
------------------------------------------------------------------------------
```

The coefficient for respiratory is equal to 2.151 indicating a positive affect as current exposure to dust increases by one unit. The column labelled "`P>|t|`" gives the p-value for the statistical test that the regression coefficients differ from zero. This tells us that the constant is significantly different from zero which is not particularly interesting as we do not expect the intercept of the line with the y-axis to be zero. It also gives a p-value of 0.012 for the test that the gradient differs from zero. There is suggestion of a positive gradient, which is significant at the conventional 5% significance level.

Alternatively STATA gives the option of reporting the odds ratio for an outcome, by clicking the odds ratio button in the reporting screen,



Or by adding the option of OR on to the end of the command

```
logit asthma respdust if group == 1, or
```

then the odds ratio is reported instead of the coefficients.

```
Iteration 0:   log likelihood = -23.064955
Iteration 1:   log likelihood = -19.864628
Iteration 2:   log likelihood = -19.134899
Iteration 3:   log likelihood = -19.118417
Iteration 4:   log likelihood =  -19.11838

Logistic regression                               Number of obs   =         73
                                                  LR chi2(1)      =       7.89
                                                  Prob > chi2     =     0.0050
Log likelihood =  -19.11838                       Pseudo R2       =     0.1711

------------------------------------------------------------------------------
      asthma | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    respdust |   8.597674   7.320396     2.53   0.012     1.620446    45.61706
------------------------------------------------------------------------------
```

Odds ratios can be interpreted as follows

- OR = 1, then there is no effect on the odds of having the outcome as the independent variable increase by one unit
- OR > 1, then the odds of having the outcome increases by a multiple of OR every time the independent variable increases by one unit
- OR < 1, then the odds of having the outcome decreases by a multiple of OR every time the independent variable increases by one unit.

In this case we can say that the probability of having **asthma** increases by a multiple of 8.6 every time **respdust** increases by 1 unit.

Model Checking

Once a logistic regression model has been produced it is important to assess how well it fits the data, one way is to use a Hosmer–Lemeshow test. This compares the observed and expected numbers of positives for different subgroups of the data, if they are similar then the model can be deemed accurate. Even so, how do we determine the number of subgroups? The most common method is to rank the subjects according to there predicted probability of a positive outcome then divide them into a number of equally sized groups the number of which then number is arbitrary, but 10 is most common. A χ^2 statistic is calculated for observed and expected numbers, if this is large then the model is not adequate.

This is performed in STATA using the post estimation command `lfit`. After a logistic regression model has been fitted as shown in the previous section the command is used to assess the models goodness of fit.

```
logit asthma respdust if group==1, or
lfit, group(10)
```

Resulting in the following output.

```
Logistic model for asthma, goodness-of-fit test

  (Table collapsed on quantiles of estimated probabilities)

       number of observations =        73
             number of groups =        10
      Hosmer-Lemeshow chi2(8) =         5.81
                  Prob > chi2 =         0.6686
```

In this case the χ^2 statistic is not significantly large with a p-value of 0.6686. This indicates that the model is accurate.

However it may be the cases that the p-value was less than 0.05 then the model would have been deemed to be non-significant which may have been caused due to a missing significant predictor variable or a significant interaction between variables.

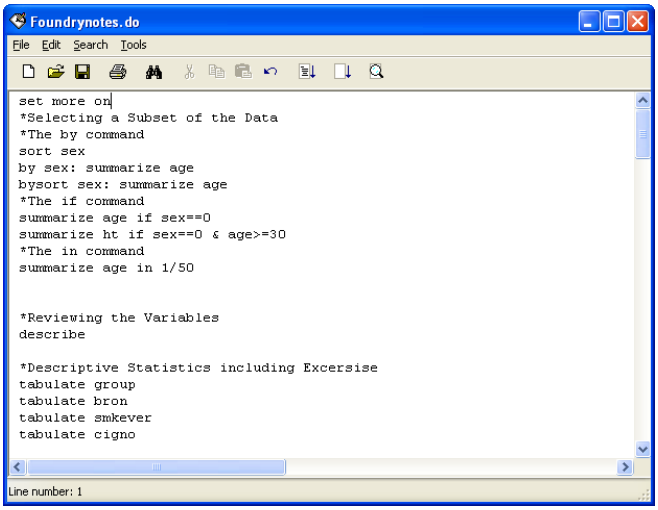
<u>Exercise</u> Examine the relationship between Bronchitis and dust levels in the exposed group using the methods above.

CREATING A STATA DO-FILES

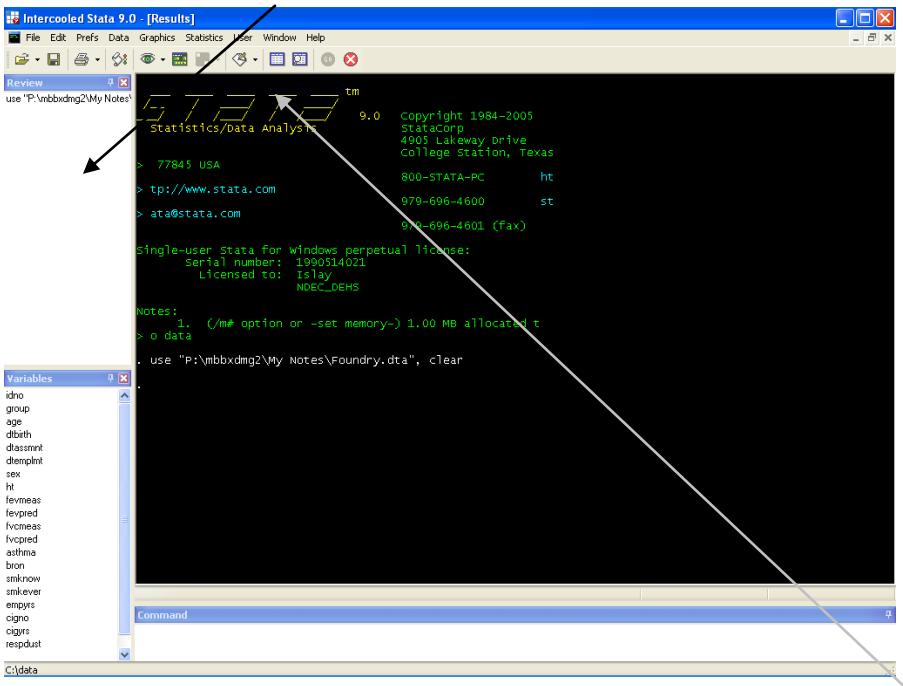
To date we have predominantly used STATA interactively through the menus, but the commands method has the advantage that it is possible to create an STATA do-file containing the commands. There are two reasons for this: -

- It makes it easier and quicker to rerun an analysis if we make changes to the raw data.
- It documents the analysis that we have performed.

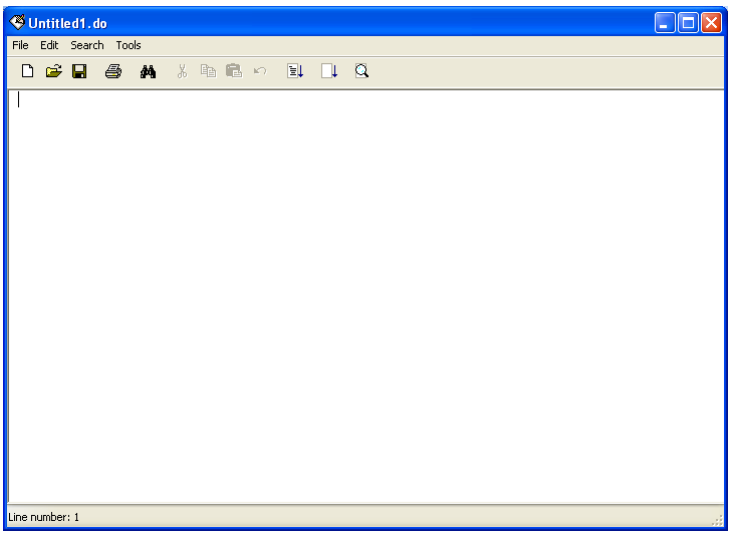
The screen shot below illustrates part of the syntax file for the analysis that we have done.



This looks complicated but STATA does make it easier for us. As previously discussed, when performing a statistical analysis using the interactive menu the corresponding STATA command is produced and logged in the review window.



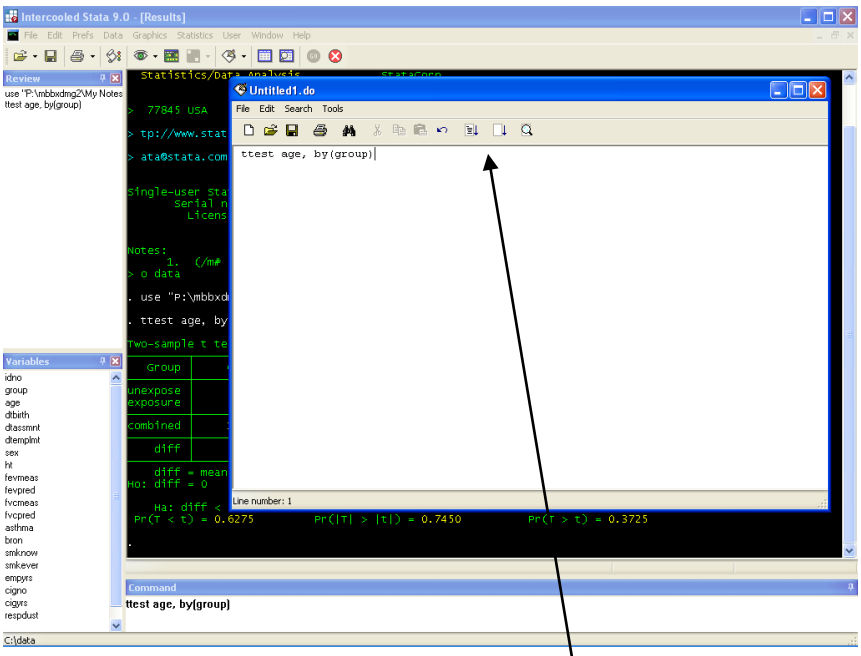
The copy and paste function can then be used to move either an individual command or all commands performed in that session into a do-file. To open a do-file simply click the **New Do-file** button located at the top of the screen. A blank do-file screen (as shown below) will then appear.



The command is either written manually or through the copying and pasting from the main STATA window.

To copy an individual command, click on the command of choice in the **review window** which will automatically place the command into the **command window**, highlight the command right click, click **copy** and then **paste** into the do-file window. Alternatively if you wish to copy all commands performed in the STATA session, right click on the review window and then click **Copy Review Contents to clipboard** and paste into the do-file.

This can be illustrated using the t-test command. Once the command has been run it can be copy and pasted into a blank do-file.



The same method as for the t-test above can be used to add further commands to the syntax. To run the entire do-file all at once simply click the **Do current file** button. Also note, it is possible to run the do-file so that it shows no output by clicking the **run current file** button located next to the **Do**

current file button. This is useful in a situation where there is a lot of output that can be ignored, in this case it is possible to tell STATA to show the output of interest by typing `noisily` in front of the corresponding command. In the t-test example above to tell STATA to show the output no matter what, the command in the do-file should be written.

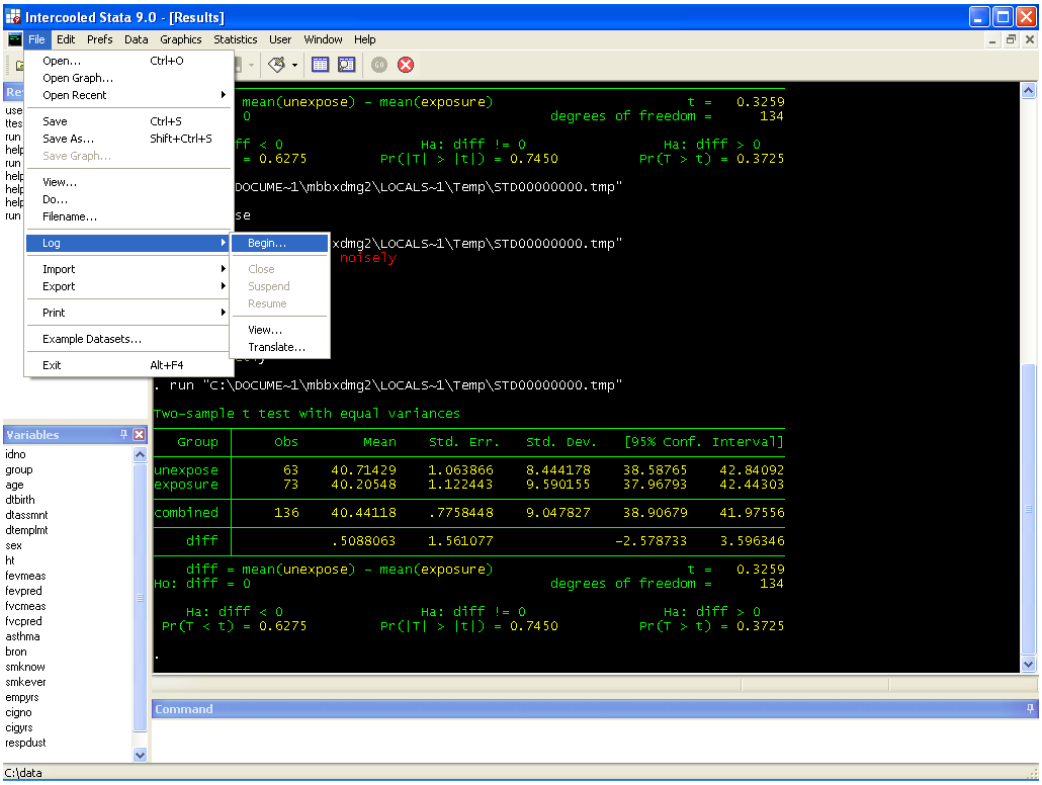
```
noisily ttest age, by(group)
```

The do-file is a separate file from STATA and needs to be saved separately at the end of the session, using **File** and **Save**. At the start of a new session, you can reopen an existing do-file, by first opening a blank do-file and then using the traditional **File**, **Open** within the do-file screen.

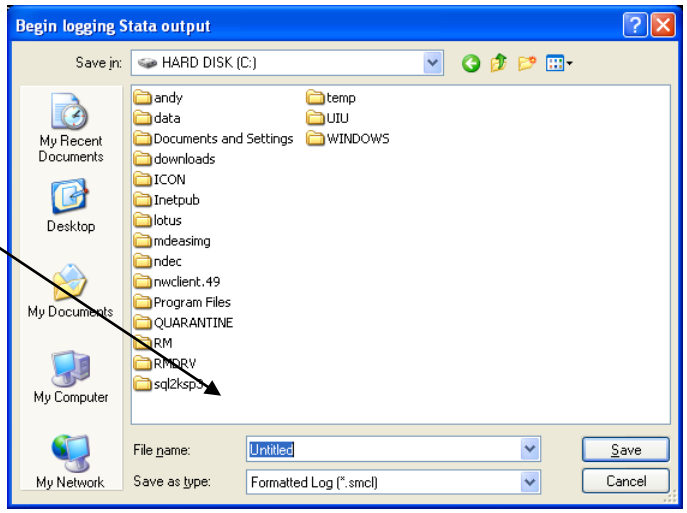
Creating a Log File

It may be the case especially when using a do-file, that you have produced a large amount of output. Unfortunately STATA will only show a certain amount in one go in its output window, hence it may not be possible to see all of your results. To avoid this problem STATA allows you to create a **Log file** which not only allows you to see all of your results but provides a method of saving the output.

The log is started by clicking **File** > **Log** > **Begin**.



STATA then asks where you would like to save the file and to give it a name.



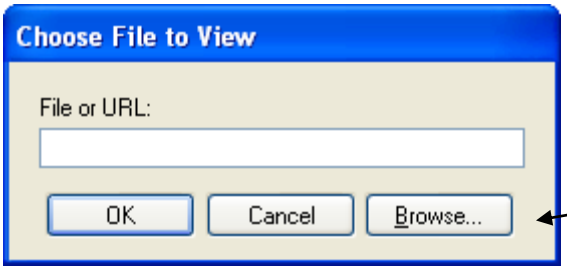
When a suitable location and name has been selected click the save button and the log file will begin.

Alternatively the command

```
log using "full file directory & name"
```

will also start the log file.

Then run the analysis you require and the once completed click **File** > **Log** > **Close** or type `log close`. Note once the Log file is started it will recorded everything that runs through the results window until the log file is closed. Also note, a log file cannot record graphs so if there is one contained in your analysis you will have to save it separately by clicking **File** > **Save Graph** or **Right clicking** on the graph followed by **Save Graph**.



To view the log file simply **Click File** > **Log** > **View…** then locate the file using the browse button and click **ok**. Alternatively, if you now the file directory route then you can type;

```
view "file directory route and name"
```

Note, the filename must end in `.smcl` for this to work.

A separate window will appear showing the contents of the log file. For example, if a log file is created for the do-file discussed in the previous section then the log file output should look like the following.

Viewer (#1) [view "P:\mbbxdmg2\My Notes\foundrylog.smcl"]

Back | Refresh | Search | Help! | Contents | What's New | News

Command: view "P:\mbbxdmg2\My Notes\foundrylog.smcl"

```
      log:  P:\mbbxdmg2\My Notes\foundrylog.smcl
 log type:  smcl
opened on:  14 Nov 2006, 15:35:30

. do "C:\DOCUME~1\mbbxdmg2\LOCALS~1\Temp\STD00000000.tmp"

. set more on

. *Selecting a Subset of the Data
. *The by command
. sort sex

. by sex: summarize age

-------------------------------------------------------------------------------
-> sex = male

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |        63    40.66667    9.863586         27         62

-------------------------------------------------------------------------------
-> sex = female

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |        73    40.24658    8.344627         24         59

. bysort sex: summarize age

-------------------------------------------------------------------------------
-> sex = male

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |        63    40.66667    9.863586         27         62

-------------------------------------------------------------------------------
```

CHOOSING THE APPROPRIATE STATISTICAL PROCEDURE

In this tutorial we have illustrated some of the basic statistical procedures available in STATA. These are summarised in the table below.

	Plausibly Continuous and Normal	**Ordinal or Ordered Categorical**	**Binary and Unordered Categories**
Comparison of Independent Two Groups	Box-plot Independent groups t-test	Box-plot or Cross-tabulation of ordered categories Mann-Whitney U-test	Cross-tabulation Chi-squared test Fisher's exact test
Comparison of more than Two groups	Analysis of variance (ANOVA)	*Kruskal Wallis analysis of Variance**	Cross-tabulation Chi-squared test
Comparison of two related outcomes	Paired samples t-test	Wilcoxon Matched Pairs	McNemar's Test
Relationship between a dependent variable and one or more independent variables	Scatter plot Regression *Pearson's correlation coefficient*	*Spearman correlation or Kendall's correlation coefficient*	*Phi coefficient*

* Not illustrated

For a more comprehensive chart for selecting methods see;
www.graphpad.com/www/book/choose.htm.